QUANTUM TRANSPORT IN ULTRASMALL DEVICES

NATO ADVANCED STUDY INSTITUTE – IL CIOCCO, ITALY

Il Ciocco

17-30 JULY

19951004 004

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	9/14/95	Final Technical 4/1/94-3/31/95

4. TITLE AND SUBTITLE
Quantum Transport in Ultrasmall Devices: NATO ASI

5. FUNDING NUMBERS
N00014-94-1-0630

6. AUTHOR(S)
David K. Ferry, Harold L. Grubin, Carlo Jacoboni and Antti-Pekka Jauho

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Arizona State University
Center for Solid State Electronics Research
Tempe, AZ 85287-6206

8. PERFORMING ORGANIZATION REPORT NUMBER
N95-5

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217-5660

DTIC
ELECTE
OCT 0 6 1995
F

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
3145030---01

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE
N00179

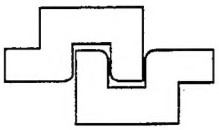
13. ABSTRACT *(Maximum 200 words)*

As a method of training new young people in the approaches that have been used, a NATO Advanced Study Institute was held at Il Ciocco, Lucca, Italy, during 17-30 July 1994, on the subject of "Quantum Transport in Ultrasmall Devices." This volume is the proceedings of that ASI. We include not only the detailed manuscripts of the major lecturers and seminar speakers, but also a series of contributed papers from the "students," themselves active researchers in this area, but who by and large are just now beginning their studies and careers.

DTIC QUALITY INSPECTED 5

14. SUBJECT TERMS	15. NUMBER OF PAGES
Quantum transport, Nanostructures, Semiconductor devices, Gate lengths	544
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL



QUANTUM TRANSPORT IN ULTRASMALL DEVICES

NATO ADVANCED STUDY INSITUTE – IL CIOCCO, ITALY

PROGRAM SCHEDULE

Sunday, July 17

3:00 p.m. Registration

8:00 p.m. Group viewing of World Cup '94 finals

Monday, July 18

9:00 a.m. "Introduction to the School," Dave Ferry
(Contained in notes beginning on p. 1)

9:30 a.m. "Introduction to Quantum Effects in Transport, Pt. 1," Carlo Jacoboni
(Notes begin on p. 1)

11:00 a.m. Coffee

11:30 a.m. "Recursive Tight-Binding Green's Function Method: Application to Ballistic and Dissipative Transport in Semiconductor Nanostructures," Fernando Sols
(Notes are not contained in this volume)

4:00 p.m. "Quantum Confined Systems: Wells, Wires, and Dots, Pt. 1," Ulrich Rössler
(Notes begin on p. 20)

5:00 p.m. Coffee

5:30 p.m. "Fabrication of Nanoscale Devices, Pt. 1," Mark Reed
(Notes begin on p. 36)

6:00 p.m. Reception Party

Accesion For
NTIS CRA&I ☑
DTIC TAB ☐
Unannounced ☐
Justification .....
By .....
Distribution /
Availability Codes
Dist | Avail and / or Special
A-1

Tuesday, July 19

9:00 a.m. "Introduction to Quantum Effects in Transport, Pt. 2," Carlo Jacoboni
(Notes begin on p. 1)

10:00 a.m. "Traditional Modeling of Semiconductor Devices, Pt. 1," Chris Snowden
(Notes begin on p. 47)

11:00 a.m. Coffee

11:30 a.m. "Effect of Band Structure and Electric Fields on Resonant Tunneling Dynamics," Gerald Iafrate
(Notes begin on p. 132)

4:00 p.m. "Quantum Confined Systems: Wells, Wires, and Dots, Pt. 2," Ulrich Rössler
(Notes begin on p. 20)

5:00 p.m. Coffee

5:30 p.m. "Fabrication of Nanoscale Devices, Pt. 2," Mark Reed
(Notes begin on p. 36)

Wednesday, July 20

9:00 a.m. "Fabrication of Nanoscale Devices, Pt. 3," Mark Reed
(Notes begin on p. 36)

10:00 a.m. "Traditional Modeling of Semiconductor Devices, Pt. 2," Chris Snowden
(Notes begin on p. 47)

11:00 a.m. Coffee

11:30 a.m. "Two-Dimensional Dynamics of Electrons Passing Through a Point Contact," Carlo Jacoboni
(Notes are not contained in this volume.)

4:00 p.m. "Quantum Confined Systems: Wells, Wires, and Dots, Pt. 3," Ulrich Rössler
(Notes begin on p. 20)

5:00 p.m. Coffee

5:30 p.m. "Mesoscopic Devices—What are They?, Pt. 1," Trevor Thornton
(Notes begin on p. 65)

Thursday, July 21

9:00 a.m. "Fabrication of Nanoscale Devices, Pt. 4," Mark Reed
(Notes begin on p. 36)

10:00 a.m. "Traditional Modeling of Semiconductor Devices, Pt. 3," Chris Snowden
(Notes begin on p. 47)

11:00 a.m. Coffee

11:30 a.m. "Artificial Impurities in Quantum Wires and Dots," Andy Sachrajda
(Notes begin on p. 140)

4:00 p.m. "Quantum Confined Systems: Wells, Wires, and Dots, Pt. 4," Ulrich Rössler
(Notes begin on p. 20)

5:00 p.m. Coffee

5:30 p.m. "Mesoscopic Devices—What are They?, Pt. 2," Trevor Thornton
(Notes begin on p. 65)

Friday, July 22

9:00 a.m. "Introduction to Quantum Effects in Transport, Pt. 2," Dave Ferry
(Notes begin on p. 1)

10:00 a.m. "Traditional Modeling of Semiconductor Devices, Pt. 4," Chris Snowden
(Notes begin on p. 47)

11:00 a.m. Coffee

Wednesday, July 27

Thursday, July 28

Friday, July 29

9:00 a.m.	"Interacting and Coherent Time-Dependent Transport in Semiconductor Heterostructures, Pt. 4," Antti-Pekka Jauho (Notes begin on p. 87)
10:00 a.m.	"Density Matrix Simulations of Semiconductor Devices, Pt. 4," Hal Grubin (Notes begin on p. 101)
11:00 a.m.	Coffee
11:30 a.m.	"Conductance in Quantum Boxes: Interference and Single Electron Effects," Mike Pepper (Notes begin on p. 180)
4:00 p.m.	"Some Recent Developments in Quantum Transport in Mesoscopic Structures and Quantum Wells, Pt. 4," Laurence Eaves (Notes begin on p. 80)
5:00 p.m.	Coffee
5:30 p.m.	"Screening and Many-Body Effects in Low-Dimensional Electron Systems, Pt. 4," Sankar Das Sarma (Notes begin on p. 122)

Closing of the School

Other Information of Interest

Meals

Breakfast	7:30-8:45 a.m.
Lunch	1:00-2:00 p.m.
Dinner	7:30- p.m.

Coffee Breaks

Morning	11:00-11:30 a.m.
Evening	5:00-5:30 p.m.

Open recreation period

Afternoons	2:00-4:00 p.m.

The Lectures and Seminar Manuscripts

The Lectures

The Seminars

INTRODUCTION TO QUANTUM EFFECTS IN TRANSPORT

C. Jacoboni[1] and D. K. Ferry[2]

[1]Dipartimento di Fisica
Universitá di Modena
41100 Modena, Italy

[2]Center for Solid State Electronics Research
Arizona State University
Tempe, Arizona 85287 USA

INTRODUCTION

Since the introduction of integrated circuits, the number of individual transistors on a single chip has doubled approximately every three years. Today, we are talking about multi-megabit DRAM memories (the 16 Mb is on the market, the 64 Mb is in pre-production, and research versions of the 256 Mb have been demonstrated) and dense signal-processing chips with comparable component density. At the rate of progress of dynamic memory (DRAM), *we can expect to reach chip densities of 10^9 devices by 2001. By 2020, we may well need to have memory chips with 1 Tb.* In general, progress in the integrated circuit field has followed a complicated scaling relationship. The reduction of design rule (or effective gate length) proceeds approximately by a factor of 1.4 each generation (which produces only an increase of 2× in density, the remainder coming from circuit enhancements and larger chip size). This means we will be using 0.1-0.15 μm rules for the 4 Gb chips (the 256 Mb chip will use 0.25 μm design rules). If we continue this extrapolation, current technology will require 30 nm design rules, and a cell size < 10^3 nm^2, for a 1 Tb memory chip.

Indeed, scaling has been followed for more than 30 years in the semiconductor industry, and it is quite clear that expectations are for it to continue at least for another decade, if not two. The American consortium SEMATECH has been working with the industries, and with universities, to prepare a new generation *roadmap* for the Semiconductor Industries Association. This roadmap illustrates a continuing linear scaling and a 0.08 μm gate length in the year 2010 (64 Gbit memories). Some projections continue the scaling of CMOS for several generations beyond this.

It appears then that we will eventually see devices with gate lengths on the order of 0.05 μm. An electron traveling at the saturated velocity (in Si) will traverse this length in 0.5 ps, or approximately the time duration of the carrier transient response at 50 kV/cm. Moreover, the inelastic mean free path for the carriers (the distance over which they lose phase information) is of the order of 0.05–0.1 μm, or comparable to the gate length (this is to distinguished from the *coherence length* or the mesoscopic *thermal length*, which are discussed in other articles of this volume). Thus, it is expected that quantum effects will certainly appear in such devices. There is evidence of this today. When devices of 30 nm (or less) gate length are made in the research laboratory, it is found that their performance is different from that of current production FETs (whose gate lengths are > 0.1 μm). Research devices fabricated both at *Arizona State University* (Ryan *et al.*, 1989; Han *et al.*, 1990), and at Sony (Ishibashi *et al.*, 1988) in Japan, with gate lengths of 25-30 nm, clearly show that tunneling through the gate depletion region is a significant contributor to current, and gate control is much reduced due to this effect (Ferry, 1990).

The transport of carriers in semiconductor devices has long been a subject of much interest, not only for material evaluation, but also in the realm of device modeling and, more importantly, as an illuminating tool for delving into the physics governing the interaction of electrons (or holes) with their environment. From the above discussion, it appears that more detailed modeling of quantum effects needs to be included in device modeling for future ultrasmall devices (Barker and Ferry, 1980a, 1980b; Ferry and Barker, 1980). These quantum effects appear in many guises: a) modification of the statistical thermodynamics, b) introduction of new length scales, c) ballistic transport and quantum interference, and d) new fluctuations affecting device performance. Many of these effects have been studied, either in models of ultra-submicron devices (or, more appropriately referred to as structures since they may well not be devices in the normal sense), or in transport studies of nanostructures (mesoscopic devices) at low temperature. The purpose of this school, and of this volume, is to review the physics, both the transport and the experimental observables, and the approaches to quantum effects in ultrasmall devices.

THE SCHOOL AND LECTURES

We bring together here two major themes: (1) the effects that can be seen in ultrasmall (mesoscopic-sized) devices, and (2) a description of the approaches that can be used to model these effects. The discussions of the experimental physics will range from the methods of fabrication of ultrasmall structures (Reed), through the physical observables in mesoscopic structures (Thornton and Eaves), to the second-order effects that can arise from fluctuations in the system (McGill). On the other hand, we begin below with an introduction to the quantum terminology for transport, along with a discussion of the waveguide theory of quantum structures. The general properties of size quantized systems will be discussed (Rössler). In addition, semi-classical approaches to transport (Snowden) will be discussed, and then followed by a general overview of quantum approaches to device modeling (Grubin). This is finally followed with discussion of the more fundamental non-equilibrium Green's's functions for modeling (Jauho) and multi-particle interactions (Das Sarma). Finally, a series of seminars will highlight recent problems and applications of quantum transport in various devices and device-like structures.

The aim is to discuss the relevant physics and transport techniques, both theoretically and experimentally, that allows the student to understand both the approaches to quantum transport in devices, and the reason why each approach is used. With this outline, it is hoped that the students for this ASI will be able to move quickly into their own research work addressing a number of the many unresolved questions. A by-product of the lectures is hopefully an awareness of these unresolved questions.

CLASSICAL AND QUANTUM DYNAMICS OF A SYSTEM OF N PARTICLES

Classical dynamics of a system of N particles

In classical mechanics the state of a system of N particles is described, in the Hamiltonian formalism, by the coordinates q_i and momenta p_i of each particle as functions of time:

$$q_i(t), p_i(t) \ . \tag{1}$$

A set of values $\{q_i, p_i\}$ represents a point in phase-space. The dynamics is governed by the Hamilton equations

$$\frac{\partial q_i(t)}{\partial t} = \frac{\partial H}{\partial p_i} \quad , \quad \frac{\partial p_i(t)}{\partial t} = -\frac{\partial H}{\partial q_i} \quad , \tag{2}$$

where H is the Hamiltonian of the system.

From the Hamilton equations, it is immediate to obtain the time variation of a quantity $u(q_i, p_i, t)$

$$\frac{du}{dt} = [u, H]_{\mathrm{P}} + \frac{\partial u}{\partial t} \ , \tag{3}$$

where $[\ldots]_{\mathrm{P}}$ indicates the Poisson braket:

$$[F, G]_{\mathrm{P}} = \sum_i \left(\frac{\partial F}{\partial q_i}\frac{\partial G}{\partial p_i} - \frac{\partial F}{\partial p_i}\frac{\partial G}{\partial q_i} \right) . \tag{4}$$

Quantum dynamics of a system of N particles

In quantum mechanics, the state of the system is described by its state vector that in the Schrödinger picture is a function of time: $|\Psi(t)\rangle$. The wavefunction

$$\Psi(q,t) = \langle q|\Psi(t)\rangle \tag{5}$$

is the probability amplitude of finding the system in q. Here q stands for all variables $\{q_i\}$ as, later, p will stand for all $\{p_i\}$. The wavefunction contains also information about momenta. In fact the wavefunction in momentum space is given by

$$\Phi(p,t) = \langle p|\Psi(t)\rangle = \int \langle p|q\rangle\, dq \langle q|\Psi(t)\rangle = \frac{1}{\sqrt{(2\pi)^{3N}}} \int e^{-ipq/\hbar} \Psi(q,t)\, dq \ . \tag{6}$$

The dynamics of the state vector in the Schrödinger picture is governed by the S.E.:

$$i\hbar \frac{d}{dt}|\Psi(t)\rangle = \mathbf{H}|\Psi(t)\rangle \ , \tag{7}$$

where, now, $\mathbf{H}$ is the Hamiltonian operator. The dynamics of the system can be described in terms of the evolution operator that applied to the state vector at time t_0 yields the state vector at time t

$$|\Psi(t)\rangle = \mathbf{U}(t, t_0)|\Psi(t_0)\rangle \ . \tag{8}$$

The differential equation for $\mathbf{U}$ is the basis dynamic law in quantum theory:

$$i\hbar \frac{d\mathbf{U}(t, t_0)}{dt} = \mathbf{H}\mathbf{U}(t, t_0) \ ; \quad \mathbf{U}(t_0, t_0) = 1. \tag{9}$$

By application of the unitary transformation $\mathbf{U}^+$ we go from the Schrödinger picture to the Heisenberg picture. The state vector in this picture is constant:

$$|\Psi_H\rangle = \mathbf{U}^+(t, t_0)|\Psi(t)\rangle = |\Psi(t_0)\rangle \ , \tag{10}$$

and the time evolution is assigned to the observables

$$\mathbf{A}_H(t) = \mathbf{U}^+(t, t_0)\mathbf{A}\mathbf{U}(t, t_0) \ . \tag{11}$$

The dynamic equation for such observables is the Heisenberg equation:

$$i\hbar \frac{d\mathbf{A}_H}{dt} = [\mathbf{A}_H, \mathbf{H}] + i\hbar \frac{\partial \mathbf{A}_H}{\partial t} \ , \tag{12}$$

where the last term accounts for a possible explicit dependence of $\mathbf{A}$ upon time (even in the Schrödinger picture).

Let us note that the classical description in terms of the values of the dynamical variables as functions of time as indicated in (1) is closer to the H. picture, where the dynamical observables are assumed to vary with time, than to the S. picture. The formal similarity between classical and quantum mechanical description of a system of N particles in the Heisenberg picture is even more evident if we write the Heisenberg equations explicitely for the q and p variables

$$\frac{d\,\mathbf{q}_H}{dt} = \frac{1}{i\hbar}[\mathbf{q}_H, \mathbf{H}] = \frac{\partial \mathbf{H}}{\partial \mathbf{p}} \ , \tag{13a}$$

$$\frac{d\,\mathbf{p}_H}{dt} = \frac{1}{i\hbar}[\mathbf{p}_H, \mathbf{H}] = -\frac{\partial \mathbf{H}}{\partial \mathbf{q}} \ . \tag{13b}$$

The main difference between these Heisenberg equations and (2) is that, here, the quantities involved are operators.

From quantum to classical dynamics

In order to obtain the classical limit from (13), we first take mean values

$$\frac{d}{dt}\langle \mathbf{q} \rangle = \left\langle \frac{\partial \mathbf{H}}{\partial \mathbf{p}} \right\rangle \ , \tag{14a}$$

$$\frac{d}{dt}\langle p\rangle = -\left\langle \frac{\partial H}{\partial q} \right\rangle \quad . \tag{14b}$$

Then, we should assume the state of the particles to be formed by wave packets small enough to allow us to neglect their dimensions with respect to any other dimensions of the physical system. At the same time the applied potential should vary at most quadratically in the region where the wavefunction is different from zero. Higher-order variations should be negligible owing to the limited region of q (for such a reason a potential step is never classical). Under the above conditions the average values on the r.h.s. of (14) can be substituted by the derivatives of the hamiltonian functions evaluated at the mean values of their arguments so that the dynamics of the mean values is described by the same equations as the classical variables.

CLASSICAL AND QUANTUM STATISTICAL MECHANICS

If a system is not completely known, as it is always the case for a system containing a large number of particles, we have to apply the methods of statistical physics in order to deal with our incomplete knowledge. The basic concept is the *statistical ensemble*, which is supposed to be known in the present notes.

The distribution function for the classical ensemble - Liouville equation

In classical mechanics a state is represented by a point in phase space: a statistical ensemble is therefore represented by a swarm of such points. The density function $D(q_i, p_i, t)$ of the ensemble indicates the density of representative points in (q_i, p_i) at time t. If we consider q_i and p_i as functions of time, then D has an implicit time dependence through q_i, p_i and an explicit dependence (a variation with time at fixed values of q_i and p_i). Its rate of change is given by (3)

$$\frac{dD}{dt} = [D,H]_P + \frac{\partial D}{\partial t} \quad . \tag{15}$$

From classical mechanics we know (Goldstein, 1959) that following a point of the ensemble along its motion, the density D of points that surround it is constant in time

$$\frac{dD}{dt} = 0 \, . \tag{16}$$

Thus, from (15) we obtain the Liouville equation

$$\frac{\partial D}{\partial t} = [H,D]_P \, . \tag{17}$$

Reduction to a single-particle description. The ensemble density D contains too detailed information: it gives us the probability of finding particle 1 in a phase-space element around (q_1,p_1), particle 2 around (q_2,p_2), and so on. In many cases it is sufficient to know the average number of particles with coordinate q between q and $q+dq$ and momentum p between p and $p+dp$. Such a quantity is given by the *single-particle distribution function*, given by

$$f(q,p,t) = \int dq_1 dp_1 \sum_i \delta(q-q_i)\delta(p-p_i)D(q_i,p_i,t) \quad . \tag{18}$$

In order to find the equation of motion for this distribution function, we differentiate with respect to time and use the Liouville equation (17) for D. This expression can be extremely complicated since it contains all interactions among the particles. If we consider identical and, as a first approximation, non interacting particles, then the Hamiltonian of the whole system is given by a sum of identical single-particle Hamiltonians:

$$H = \sum_i H^{(sp)}(q_i,p_i); \qquad \dot{q}_i = \frac{\partial H^{(sp)}(q_i,p_i)}{\partial p_i} \quad ; \qquad \dot{p}_i = -\frac{\partial H^{(sp)}(q_i,p_i)}{\partial q_i} \tag{19}$$

and

$$\frac{\partial D}{\partial t} = [H,D]_P = \sum_i \left(\frac{\partial H}{\partial q_i}\frac{\partial D}{\partial p_i} - \frac{\partial H}{\partial p_i}\frac{\partial D}{\partial q_i} \right) = -\sum_i \left(\frac{\partial D}{\partial q_i}\dot{q}_i + \frac{\partial D}{\partial p_i}\dot{p}_i \right) \, . \tag{20}$$

If this result is used in the time derivative of (18), the result

$$\frac{\partial}{\partial t} f(q,p,t) = -\int dq_1 dp_1 \sum_i \delta(q-q_i)\delta(p-p_i) \sum_j \left(\frac{\partial D}{\partial q_j}\dot{q}_j + \frac{\partial D}{\partial p_j}\dot{p}_j \right) \tag{21}$$

can be integrated by parts, and with the help of some regularity conditions on D, the following result is obtained (Bogoliubov and Bogoliubov, 1982; Balescu, 1975)

$$\frac{\partial}{\partial t} f(q,p,t) + \dot{q}\frac{\partial}{\partial q} f + \dot{p}\frac{\partial}{\partial p} f = 0 \quad , \tag{22}$$

i.e., the *Boltzmann equation* (BE) for free identical particles.

If the interaction among particles has to be introduced, a collision integral will be added to the BE:

$$\frac{\partial}{\partial t} f(q,p,t) + \dot{q}\frac{\partial}{\partial q} f + \dot{p}\frac{\partial}{\partial p} f = \int dp' \left[f(q,p',t)P(q,p',p) - f(q,p,t)P(q,p,p') \right] \, , \tag{23}$$

where $P(q,p,p')$ represents the probability per unit time that a particle in q will be scattered from p to p'. In general, P will contain a two-particle distribution function $f(r_1,p_1,r_2,p_2)$, proportional to the joint probability of having a particle in r_1 with p_1 and a particle in r_2 with p_2. A three-particle distribution will enter the equation for $f(r_1,p_1,r_2,p_2)$, and so on. A hierarchy of equations is thus obtained that must be truncated with some approximation, for example by assuming the the two-particle distribution is given by the product of two one-particle distributions. This would mean to use the independent particle approximation for the evaluation of their collisions.

Other scattering mechanisms are often inserted in the collision integral (impurities, phonons) and collisions are in general assumed to be instantaneous in time and point-like in space. Most of the approximations necessary to apply classical or semiclassical statistical physics fail, or at least lack justification, when applied to nanoelectronic devices.

The density matrix for the quantum ensemble - Liouville-von Neumann equation

In quantum mechanics a point in phase space contradicts the uncertainty principle since it would correspond to a well defined set of positions and momenta. Statistical physics is still based, however, on the concept of the statistical ensemble. Given a physical system, described by the state vector $|\Psi\rangle$, the expectation value of a measurement of a given quantity A is given by

$$\langle A\rangle = \langle\Psi|\mathbf{A}|\Psi\rangle \ , \tag{24}$$

where **A** is the operator related to the physical quantity of interest. On the other hand, if the state of the system is not completely specified, the expectation value, given in Eq. (24), is replaced by its *statistical ensemble* average:

$$\overline{\langle A\rangle} = \overline{\langle\Psi|\mathbf{A}|\Psi\rangle} \ , \tag{25}$$

where the overbar indicates an average to be performed over a suitable statistical ensemble that accounts for our partial knowledge of the system. If $\{|\alpha\rangle\}$ is a complete set of basis vectors, the above equation can also be written as

$$\overline{\langle A\rangle} = \sum_{\alpha,\alpha'}\overline{\langle\Psi|\alpha\rangle\langle\alpha|\mathbf{A}|\alpha'\rangle\langle\alpha'|\Psi\rangle} = Tr\left(\overline{|\Psi\rangle\langle\Psi|}\mathbf{A}\right) \ . \tag{26}$$

From the above result it is easy to recognize that, in quantum statistical physics, the mathematical instrument for the evaluation of average quantities is the *density-matrix operator*

$$\rho = \overline{|\Psi\rangle\langle\Psi|} \ . \tag{27}$$

By means of this operator, the direct link in (26) between quantum theory and experiments can be easily expressed as

$$\overline{\langle A\rangle} = Tr(\rho\mathbf{A}) \ , \tag{28}$$

which can be regarded as the basic equation of quantum statistical mechanics.

In a given representation $\{|\alpha\rangle\}$ the matrix elements of ρ constitute the *density matrix* of the system (Ter Haar, 1961). It is easy to show that the diagonal elements of the density matrix $\rho_{\alpha\alpha}$ give the probabilities P_α of finding the system in the various states $|\alpha\rangle$. In fact, if $P^{(i)} = 1/N$ is the probability of selecting at random the ith system, and $P(\alpha|i) = |\langle\alpha|\Psi^{(i)}\rangle|$ the probability of finding the ith system in the state α, we have

$$P_\alpha = \sum_i P^{(i)}P(\alpha|i) = \sum_i \frac{1}{N}\langle\alpha|\Psi^{(i)}\rangle\langle\Psi^{(i)}|\alpha\rangle = \left\langle\alpha\left|\overline{|\Psi\rangle\langle\Psi|}\right|\alpha\right\rangle = \rho_{\alpha\alpha} \ . \tag{29}$$

When dealing with a transport problem, we are interested in studying the time evolution of the average quantities in (28). This can be obtained starting from the time evolution of the state vector in the Schrödinger picture, given by (8). The result is

$$\rho(t) = \mathbf{U}(t,t_0)\rho(t_0)\mathbf{U}^\dagger(t,t_0) \ . \tag{30}$$

By differentiating with respect to time, we obtain the Liouville-von Neumann equation for the density matrix in the Schrödinger picture:

$$i\hbar\frac{\partial}{\partial t}\rho(t) = [\mathbf{H},\rho(t)] = \mathrm{L}[\rho(t)] \ . \tag{31}$$

Here L is the Liouvillian, a special operator that, once applied to an operator **A**, yields the commutator of the Hamiltonian with **A**. Since L acts on operators and gives a new operator as a result, it is often called a *superoperator*.

Let us now consider the equation of motion of the density-matrix operator in other pictures of quantum mechanics. In the Heisenberg picture the density matrix does not depend on time, since the state vectors are constant. If the total Hamiltonian is split into two parts ($\mathbf{H_0}$+$\mathbf{H'}$), where $\mathbf{H'}$ is considered as a perturbation, the interaction picture can be used, and the time evolution of the state vector is given by

$$|\Psi_I(t)\rangle = \mathbf{S}(t,t_0)|\Psi_I(t_0)\rangle, \tag{32}$$

where the evolution operator **S** for the state vectors in the interaction picture verifies the following differential equation:

$$i\hbar\frac{\partial}{\partial t}\mathbf{S}(t,t_0) = \mathbf{H'}_I(t)\mathbf{S}(t,t_0). \tag{33}$$

Starting from (27) and following the same theoretical development used to derive (31), we obtain the Liouville-von Neumann equation in the interaction picture:

$$i\hbar\frac{\partial}{\partial t}\rho_I(t) = [\mathbf{H'}_I(t),\rho_I(t)]. \tag{34}$$

The above equation of motion is more convenient in that only the pertubation Hamiltonian appears explicitely in the commutator.

The time evolution of an average quantity $\overline{\langle A\rangle}$, which is always given by (28), is due to the time evolution of ρ in the Schrödinger picture and to the time evolution of A in the Heisenberg picture. In the interaction picture, A carries the time evolution due to the known, unperturbed Hamiltonian, and ρ the time evolution due to the perturbation Hamiltonian.

The reduced density-matrix operator. In order to discuss the reduction problem for a given subsystem of interest, let us consider again the example of n electrons interacting with phonons in a crystal. The state vectors, and therefore the density matrix, will be functions of the electron coordinates $\mathbf{x}\equiv(\mathbf{r}_1,\mathbf{r}_1,\ldots,\mathbf{r}_N)$ and the phonon variables ξ:

$$\rho = \rho(\mathbf{x},\xi,\mathbf{x}',\xi'). \tag{35}$$

If an observable $\mathbf{A}^{(el)}$ acts only on the electron variables, then $\mathbf{A}^{(el)}(\mathbf{x},\xi,\mathbf{x}',\xi') = \mathbf{A}^{(el)}(\mathbf{x},\mathbf{x}')\,\delta(\xi-\xi')$, and its mean value is given by

$$\overline{\left\langle\mathbf{A}^{(el)}\right\rangle} = Tr(\rho\mathbf{A}^{(el)}) = \int d\mathbf{x}\int d\mathbf{x}'\int d\xi\rho(\mathbf{x},\xi,\mathbf{x}',\xi)\mathbf{A}^{(el)}(\mathbf{x}',\mathbf{x}) = Tr(\rho^{(el)}\mathbf{A}^{(el)}), \tag{36}$$

where

$$\rho^{(el)}(\mathbf{x},\mathbf{x}') \equiv \int d\xi \rho(\mathbf{x},\xi,\mathbf{x}',\xi) = Tr_{\xi}(\rho) \tag{37}$$

is the *reduced electronic density matrix.*

Reduction to a single-particle description. In practice, the observable $\mathbf{A}^{(el)}$ is often the average over the particles of the system of a single-particle quantity $\mathbf{A}^{(sp)}$. This is the case, for exemple, of the drift velocity or the mean energy of the electron gas. In such a situation, owing to the symmetry of the wave functions of identical particles, and therefore of the corresponding density matrix, the average $\langle \mathbf{A}^{(el)} \rangle$ can be expressed as

$$\begin{aligned} \overline{\langle \mathbf{A}^{(el)} \rangle} &= \int d\mathbf{x} \int d\mathbf{x}' \, \rho^{(el)}(\mathbf{x},\mathbf{x}') \frac{1}{N} \sum_i \mathbf{A}^{(sp)}(\mathbf{r}_i', \mathbf{r}_i) \\ &= \int d\mathbf{r}_1 \ldots \int d\mathbf{r}_N \int d\mathbf{r}'_1 \ldots \int d\mathbf{r}'_N \times \rho^{(el)}(\mathbf{r}_1, \ldots \mathbf{r}_N; \mathbf{r}'_1, \ldots, \mathbf{r}'_N) \mathbf{A}^{(sp)}(\mathbf{r}'_1, \mathbf{r}_1) \quad , \end{aligned} \tag{38}$$

which is the quantum analogue of the classical reduction in (18). Such a result suggests the introduction of a single-particle density matrix $\rho^{(sp)}$ as the integral of the reduced electronic density matrix in (37) over all the coordinates but one:

$$\rho^{(sp)}(\mathbf{r},\mathbf{r}') \equiv \int d\mathbf{r}_2 \ldots \int d\mathbf{r}_N \int d\mathbf{r}'_2 \ldots \int d\mathbf{r}'_N \times \rho^{(el)}(\mathbf{r},\mathbf{r}_2, \ldots \mathbf{r}_N; \mathbf{r}', \mathbf{r}'_2, \ldots, \mathbf{r}'_N). \tag{39}$$

The average value is then given by

$$\overline{\langle A^{(el)} \rangle} = \int d\mathbf{r} \int d\mathbf{r}' \rho^{(sp)}(\mathbf{r},\mathbf{r}') \mathbf{A}^{(el)}(\mathbf{r}',\mathbf{r}) = Tr(\rho^{(sp)} \mathbf{A}^{(sp)}). \tag{40}$$

The above two reduction processes can be introduced in a more general and elegent way in the second-quantization formalism. If the observable of interest is a single-electron quantity, the corresponding single-particle operator in second quantization is given by

$$\int d\mathbf{r} \int d\mathbf{r}' \, \Psi^+(\mathbf{r}') \mathbf{A}^{(el)}(\mathbf{r}',\mathbf{r}) \Psi(\mathbf{r}), \tag{41}$$

where $\Psi^+(\mathbf{r})$ and $\Psi(\mathbf{r})$ are the second-quantized creation and annihilation operators, respectively. The average value is

$$\overline{\langle A^{(el)} \rangle} = \int d\mathbf{r} \int d\mathbf{r}' \overline{\langle \Psi^+(\mathbf{r}') \Psi(\mathbf{r}) \rangle} A^{(el)}(\mathbf{r}',\mathbf{r}). \tag{42}$$

From the comparison of this expression with (40), we obtain the form of the single-particle density matrix in second quantization:

$$\rho^{(sp)}(\mathbf{r},\mathbf{r}') = \overline{\langle \Psi^+(\mathbf{r}') \Psi(\mathbf{r}) \rangle} = Tr\{\rho \Psi^+(\mathbf{r}') \Psi(\mathbf{r})\}. \tag{43}$$

This second-quantization approach is more general in that it allows one to investigate physical systems with a variable number of particles.

The single-particle density matrix is often the quantity we need in order to describe our transport problem. However, if we try to write down an equation of motion for $\rho^{(sp)}$ starting from the Liouville-von Neumann equation, we obtain from (31) and (43)

$$i\hbar \frac{\partial}{\partial t} \rho^{(sp)}(\mathbf{r},\mathbf{r}') = Tr\{[\mathbf{H},\rho] \Psi^+(\mathbf{r}') \Psi(\mathbf{r})\}. \tag{44}$$

Since the Hamiltonian in general contains terms corresponding to electron-electron and electron-phonon interaction, the trace operation does not commute with $\mathbf{H}$, and we do not obtain a closed equation for $\rho^{(sp)}$. In particular, using the explicit form of $\mathbf{H}$ in terms of second quantization, it can be shown that the equation of motion for the single-particle density matrix contains the two-particle density matrix, the equation of motion for the two-particle density matrix contains the three-particle one, and so on. As for the analogous classical case, in order to obtain a closed equation of motion, a truncation in this infinitive hierarchy of equations is required. This is generally obtained replacing the average values of products of pairs of field operators with the product of their average values (mean-field approximation) (Kadanoff and Baym, 1962).

The Wigner function

The Wigner function was introduced as an extension of the concept of distribution to the quantum case, and it constitutes the more direct link between the quantum density matrix and the "classical" description of the evolution of the system in phase space through a distribution $f(\mathbf{r},\mathbf{k},t)$. If we consider, for simplicity, a single-particle system with canonical variables $\mathbf{r}$ and $\mathbf{k}$, it has been already mentioned above that we cannot define a probability function $P(\mathbf{r},\mathbf{k})$ such that $P(\mathbf{r},\mathbf{k})d\mathbf{r}\,d\mathbf{k}$ is equal to the probability of finding the system in $d\mathbf{r}$ around $\mathbf{r}$ and $d\mathbf{k}$ around $\mathbf{k}$, since this probability is ill defined in quantum mechanics owing to the incompatibility of the two necessary measurements.

The Wigner function is defined as the Weyl-Wigner transform of the density operator ρ as follows:

$$f_w(\mathbf{r},\mathbf{k},t) = \frac{1}{(2\pi)^3} \int d\mathbf{r}' e^{-i\mathbf{k}\mathbf{r}'} \left\langle \mathbf{r} + \frac{\mathbf{r}'}{2} \middle| \rho(t) \middle| \mathbf{r} - \frac{\mathbf{r}'}{2} \right\rangle . \tag{45}$$

Several interesting properties of f_w suggest thinking of this function as a quantum extension of the concept of distribution function (Hillary *et al.*, 1984; Iafrate, 1988). In particular,

$$\int f_w(\mathbf{r},\mathbf{k},t) d\mathbf{k} = \langle \mathbf{r} | \rho(t) | \mathbf{r} \rangle = \rho(\mathbf{r},\mathbf{r},t) \tag{46}$$

(position probability density), and

$$\int f_w(\mathbf{r},\mathbf{k},t) d\mathbf{r} = \langle \mathbf{k} | \rho(t) | \mathbf{k} \rangle = \rho(\mathbf{k},\mathbf{k},t) \tag{47}$$

(momentum probability density). For an observable $\mathbf{A}(\mathbf{r},\mathbf{k},t)$ it can be shown that

$$\overline{\langle A \rangle}(t) = Tr(\rho \mathbf{A}) = \frac{1}{(2\pi)^3} \int d\mathbf{r} \int d\mathbf{k} \, A_w(\mathbf{r},\mathbf{k},t) f_w(\mathbf{r},\mathbf{k},t), \tag{48}$$

where $A_w(\mathbf{r},\mathbf{k},t)$ is a function obtained from the operator **A** by applying Weyl's rule. It is not possible however to attribute a direct probabilistic interpretation to f_w, since, in general, it is not positive definite.

A physical insight into the meaning of the Wigner function can be obtained by observing that its value in a region around $(\mathbf{r},\mathbf{k})$ gives information on the $\mathbf{k}$ Fourier components of the autocorrelation of the wavefunction around $\mathbf{r}$. Thus, f_w has significant values, in general, in regions of phase space where the presence of the particles can be detected within the uncertainty principle. Some more information related to the quantum-mechanical phases are added, which can result in negative values for f_w. Furthermore, it must be noted that, since a correlation function at distant points appears in the definition of f_w, in special cases, like the Aharonov-Bohm geometry, f_w can be different from zero also in points where the particle cannot be found.

If the interaction of the particle with the rest of the system can be described by a potential V(r), an equation for the Wigner function can be written (Meijer, 1966)

$$\frac{\partial f_w(\mathbf{r},\mathbf{k},t)}{\partial t}+\frac{\hbar\mathbf{k}}{m}\frac{\partial f_w(\mathbf{r},\mathbf{k},t)}{\partial \mathbf{r}}=\int d\mathbf{k}'\, W(\mathbf{r},\mathbf{k}-\mathbf{k}')f_w(\mathbf{r},\mathbf{k}',t), \tag{49}$$

where

$$W(\mathbf{r},\mathbf{k})=\frac{1}{(2\pi\hbar)^3}\int d\mathbf{r}'\, e^{i\mathbf{k}\mathbf{r}'}\left[V(\mathbf{r}-\frac{1}{2}\mathbf{r}')-V(\mathbf{r}+\frac{1}{2}\mathbf{r}')\right]. \tag{50}$$

If $V(\mathbf{r})$ is an analytic function, its values in the integrand can be obtained with a series expansion around r, and a second form is obtained for the Wigner equation,

$$\frac{\partial f_w(\mathbf{r},\mathbf{k},t)}{\partial t}+\frac{\hbar\mathbf{k}}{m}\frac{\partial f_w(\mathbf{r},\mathbf{k},t)}{\partial \mathbf{r}}+F\frac{\partial f_w(\mathbf{r},\mathbf{k},t)}{\hbar\partial \mathbf{k}}=\sum_{\lambda=3,5,\ldots}\frac{(\hbar/2i)^{(\lambda-1)}}{\lambda!}\frac{d^\lambda V(\mathbf{r})}{d\mathbf{r}^\lambda}\frac{\partial^\lambda f_w}{\partial(\hbar\mathbf{k})^\lambda}, \tag{51}$$

which clearly reduces to the classical Boltzmann equation for $\hbar\to 0$.

A dissipative term has been sometimes added to the Wigner equation (49) in the for of a relaxation time. However it is possible also to include the electron-phonon interaction in a rigorous way (Rossi *et al.*, 1994) as briefly indicated in the following.

Wigner equation with phonon scattering. In order to extend the above theory to the case where phonon scattering is present, we consider the set of basis vectors $|\mathbf{k},\{n_q\}\rangle\equiv|\mathbf{k}\rangle|\{n_q\}\rangle$ given by the product of electron momentum eigenstates $|\mathbf{k}\rangle$ and phonon number eigenstates $|\{n_q\}\rangle$. In terms of such basis vectors we introduce the Weyl-Wigner transform of the density matrix operator over the electron coordinates

$$f_w(\mathbf{r},\mathbf{k};\{n_q\},\{n'_q\};t)=\frac{1}{(2\pi)^3}\int d\mathbf{k}' e^{i\mathbf{k}'\mathbf{r}}\left\langle \mathbf{k}+\frac{\mathbf{k}'}{2},\{n_q\}|\rho(t)|\mathbf{k}-\frac{\mathbf{k}'}{2},\{n'_q\}\right\rangle. \tag{52}$$

The conventional Wigner function will be obtained as the trace over the phonon coordinates of the above generalized Wigner function.

We can now move to the interaction picture taking as unperturbed Hamiltonian only the free-phonon term. The density matrix operator in (52) is replaced by the same operator in this interaction picture, ρ_I, and the corresponding f_{WI} will be obtained. Using the Liouville - von Neumann equation for ρ_I into the time derivative of the resulting expression, the following equation of motion for f_{WI} results (Rossi *et al.*, 1994)

$$\left(\frac{\partial}{\partial t}+\frac{\hbar\mathbf{k}}{m}\cdot\nabla\right)f_{wI}(\mathbf{r},\mathbf{k};\{n_q\},\{n'_q\};t)=\sum_{q,\{n''_q\}}e^{i\mathbf{k}\mathbf{r}}\begin{bmatrix}H_2\left(g;\{n_q\},\{n''_q\},t\right)f_{wI}\left(\mathbf{r},\mathbf{k}-\frac{\mathbf{g}}{2},\{n''_q\},\{n'_q\},t\right)\\-f_{wI}\left(\mathbf{r},\mathbf{k}+\frac{\mathbf{g}}{2},\{n_q\},\{n''_q\},t\right)H_2\left(g;\{n''_q\},\{n'_q\},t\right)\end{bmatrix} \tag{53}$$

where H_2 (...) are the matrix elements of the electron and electron-phonon Hamiltonian divided by $ih/2\pi$. Once again the trace over the phonon coordinates does not commute with the Hamiltonian and a close equation for the electron fw is not obtained. We shall see that a Monte Carlo technique can be introduced that, in principle, overcome this difficulty.

GREEN'S FUNCTIONS

The basic idea of Green's Function

The general concept of Green's function is that of a quantity that indicates how a "cause," or "source," in $\mathbf{r}'$ at time t' determines an "effect" in $\mathbf{r}$ at time t. In the case of a physical field the "causes" are given, in general, by the field sources and initial and boundary conditions. In principle, therefore, we have one Green's function for the sources and one for the initial and boundary conditions (Morse and Feshbach, 1953). In general, however, initial and boundary conditions can be treated as special sources so that only one type of Green's function need to be considered.

Let us assume that the equation satisfied by the function $f(\mathbf{r},t)$ of interest is

$$\mathbf{A}f(\mathbf{r},t)=s(\mathbf{r},t), \tag{54}$$

where **A** is some linear operator and $s(\mathbf{r},t)$ represents the source. Then the general idea of Green's function introduced above leads to the definition of a Green's function $G(\mathbf{r},t,\mathbf{r}',t')$ such that

$$f(\mathbf{r},t)=\int d\mathbf{r}'\,dt'\,G(\mathbf{r},t,\mathbf{r}',t')\,s(\mathbf{r}',t'). \tag{55}$$

Then, if G satisfies the equation

$$\mathbf{A}\,G(\mathbf{r},t,\mathbf{r}',t')=\delta(\mathbf{r}-\mathbf{r}')\delta(t-t'), \tag{56}$$

the expression in (55) is a good solution of (54). The solution of (54), however, is not unique since any solution of the homogeneous equation can be added, and this freedom can be used to satisfy the initial (and boundary) conditions at t_i.

For the cases of interest to us, where we deal with distribution functions and wavefunctions, the equations are homogeneous and the sources are given by the fields themselves. The source term in the starting equation (54) is a flash of the initial field:

$$\mathbf{A}\,f(\mathbf{r},t)=f(\mathbf{r},t_i)\delta(t-t_i), \tag{57}$$

and (55) becomes

$$f(\mathbf{r},t) = \int d\mathbf{r}'\, G(\mathbf{r},t,\mathbf{r}',t_i)\, f(\mathbf{r}',t_i) \ . \tag{58}$$

In order for f to satisfy the initial condition, G must verify the condition

$$\lim_{t\to t'} G(\mathbf{r},t,\mathbf{r}',t') = \delta(\mathbf{r}-\mathbf{r}') \ , \tag{59}$$

which is imposed as initial condition on (56).

Green's function in classical mechanics

Let us consider a distribution of particles satisfying the collisionless B.E. with a force field $F(r) = ma(r)$

$$\mathsf{L}f = \frac{\partial f}{\partial t} + \boldsymbol{v}\frac{\partial f}{\partial \boldsymbol{r}} + \boldsymbol{a}(\boldsymbol{r})\frac{\partial f}{\partial \boldsymbol{v}} = 0 \ . \tag{60}$$

Given r, v, and t, let us define $r^*(r,v,t\text{-}t_i)$ and $v^*(r,v,t\text{-}t_i)$ as the velocity of a particle that will be in r with v $t\text{-}t_i$ seconds later. It is clear that r^* and v^* are the initial conditions of the Newton trajectory (r,v,t) of interest. If we continue this trajectory for a time dt the corresponding initial conditions will not be changed:

$$\boldsymbol{r}^*(\boldsymbol{r}+\boldsymbol{v}dt,\boldsymbol{v}+\boldsymbol{a}dt,t+dt-t_i) = \boldsymbol{r}^*(\boldsymbol{r},\boldsymbol{v},t-t_i) \ , \tag{61}$$

$$\boldsymbol{v}^*(\boldsymbol{r}+\boldsymbol{v}dt,\boldsymbol{v}+\boldsymbol{a}dt,t+dt-t_i) = \boldsymbol{v}^*(\boldsymbol{r},\boldsymbol{v},t-t_i) \ , \tag{62}$$

which means that

$$\mathsf{L}\boldsymbol{r}^* = \frac{\partial \boldsymbol{r}^*}{\partial t} + \boldsymbol{v}\frac{\partial \boldsymbol{r}^*}{\partial \boldsymbol{r}} + \boldsymbol{a}(\boldsymbol{r})\frac{\partial \boldsymbol{r}^*}{\partial \boldsymbol{v}} = 0 \ ; \qquad \mathsf{L}\boldsymbol{v}^* = \frac{\partial \boldsymbol{v}^*}{\partial t} + \boldsymbol{v}\frac{\partial \boldsymbol{v}^*}{\partial \boldsymbol{r}} + \boldsymbol{a}(\boldsymbol{r})\frac{\partial \boldsymbol{v}^*}{\partial \boldsymbol{v}} = 0 \ . \tag{63}$$

Since a particle will be in r at time t with velocity v if at the initial time t_i it was in r^* with velocity v^*, and the density around a particle in phase space is constant in time, we expect the Green's function G_B for our BE equation (60) to be given by

$$G_B(\mathbf{r},\boldsymbol{v},t,\mathbf{r}',\boldsymbol{v}',t') = \delta(\mathbf{r}'-\mathbf{r}^*(\mathbf{r},\boldsymbol{v},t-t'))\delta(\boldsymbol{v}'-\boldsymbol{v}^*(\mathbf{r},\boldsymbol{v},t-t'))\theta(t-t') \ , \tag{64}$$

where the step function θ has been inserted in order to consider retarded solutions. By means of (63) it is easy to show that this function verifies the equation

$$\mathsf{L}G_B = \frac{\partial G_B}{\partial t} + \boldsymbol{v}\frac{\partial G_B}{\partial \boldsymbol{r}} + \boldsymbol{a}(\boldsymbol{r})\frac{\partial G_B}{\partial \boldsymbol{v}} = \delta(\mathbf{r}'-\mathbf{r}^*(\mathbf{r},\boldsymbol{v},t-t'))\delta(\boldsymbol{v}'-\boldsymbol{v}^*(\mathbf{r},\boldsymbol{v},t-t'))\delta(t-t'). \tag{65}$$

since r^* and v^* reduce to r and v for $t = t'$, owing to the last δ, this is equivalent to

$$\mathsf{L}G_B = \delta(\mathbf{r}-\mathbf{r}')\delta(\boldsymbol{v}-\boldsymbol{v}')\delta(t-t') \tag{66}$$

in agreement with the general Green's-function equation (56). Furthermore the limit for $t\to t'$ of (64) agrees with what required by (59).

According to our introduction, we now assume that our "source" is a "flash" distribution at time t_i:

$$s(\mathbf{r},\boldsymbol{v},t) = f(\mathbf{r},\boldsymbol{v},t_i)\delta(t-t_i), \tag{67}$$

and with the Green's function (64) we obtain the solution

$$\begin{aligned} f(\mathbf{r},\boldsymbol{v},t) &= \int d\mathbf{r}'\, d\boldsymbol{v}'\, dt'\; G_B(\mathbf{r},\boldsymbol{v},t,\mathbf{r}',\boldsymbol{v}',t')\, s(\mathbf{r}',\boldsymbol{v}',t') \\ &= \int d\mathbf{r}'\, d\boldsymbol{v}'\, dt'\; \delta(\mathbf{r}'-\mathbf{r}^*(\mathbf{r},\boldsymbol{v},t-t'))\,\delta(\boldsymbol{v}'-\boldsymbol{v}^*(\mathbf{r},\boldsymbol{v},t-t'))\theta(t-t')\, f(\mathbf{r}',\boldsymbol{v}',t_i)\delta(t'-t_i) \\ &= f(\mathbf{r}^*(\mathbf{r},\boldsymbol{v},t-t_i),\boldsymbol{v}^*(\mathbf{r},\boldsymbol{v},t-t_i),t_i) \end{aligned} \tag{68}$$

as expected for ballistic transport.

Green's Functions in quantum mechanics

In the case of the Schroedinger equation (SE), the "source" has to be considered again as a flash of wavefunction at the initial time:

$$s(q,t) = i\hbar\Psi(q,t_i)\delta(t-t_i), \tag{69}$$

where q stands for the set of variables of the system, and the factor $i\hbar$ is due to the fact that the time derivative of the differential operator of the SE contains the same factor. According to the general theory, the retarded Green's function for the SE, $G_S^{(r)}$, will be such that

$$\begin{aligned} \Psi(q,t) &= \int dq\, dt\; G_S^{(r)}(q,t,q',t')\, i\hbar\Psi(q',t_i)\delta(t'-t_i) \\ &= \int dq\; G_S^{(r)}(q,t,q',t_i)\, i\hbar\Psi(q',t_i) \end{aligned} \tag{70}$$

If we compare this expression with the definition of the evolution operator in Eq. (8), we realize that the Green's function above is the q representation of the evolution operator:

$$G_S^{(r)}(q,t,q',t') = \frac{\theta(t-t')}{i\hbar}\langle q|\mathsf{U}(t,t')|q'\rangle \ . \tag{71}$$

We may thus define a corresponding *retarded Green's operator* as

$$\mathbf{G}^{(r)}(t,t') = \frac{\theta(t-t')}{i\hbar}\mathsf{U}(t,t') \ . \tag{72}$$

The differential equation satisfied by $\mathbf{G}^{(r)}$ can be immediately obtained from the dynamical equation for the evolution operator (9):

$$\left\{i\hbar\frac{\partial}{\partial t} - \mathsf{H}\right\}\mathbf{G}^{(r)}(t,t') = \delta(t-t')\mathbb{1}. \tag{73}$$

This is coherent with (56) because the identity operator has a q-representation given by a δ function.

A similar advanced Green's operator can be defined as

$$G^{(a)}(t,t') = \frac{\theta(t'-t)}{i\hbar} U(t,t') \,, \tag{74}$$

which verifies the same Green's equation (73) and, with an equation similar to (70), yelds the state of the system at a time t previous to the initial time t_i.

The retarded and advanced Green's operators carry information on the dynamics, but not on the state of the system, as one can see from their definitions and the differential equation (73) that they satisfy. As we shall see shortly, other Green's functions can be defined that also carry information on the state of the system.

Green's Functions in second quantization

If we consider a single particle, in terms of second quantization we may write the r matrix elements of the evolution operator as

$$U(\mathbf{r},t;\mathbf{r}',t') \equiv \langle \mathbf{r}|U(t,t')|\mathbf{r}'\rangle = \langle 0|\Psi(\mathbf{r})U(t,t')\Psi^+(\mathbf{r}')|0\rangle, \tag{75}$$

where $|0\rangle$ is the vacuum state. In the Heisenberg picture the last scalar product is $\langle 0|\Psi(\mathbf{r},t)\Psi^+(\mathbf{r}',t')|0\rangle$, so that the retarded Green's function for a system containing one particle can be written as

$$G^{(r)}(\mathbf{r},t;\mathbf{r}',t') = \frac{I}{i\hbar}\langle 0|\Psi(\mathbf{r},t)\Psi^+(\mathbf{r}',t')|0\rangle. \tag{76}$$

A physical interpretation of the above expression can be given as follows: after a particle has been created at $\mathbf{r}'$ at time t' the probability amplitude is evaluated of finding the particle at $\mathbf{r}$ at time t. This interpretation suggests an alternative expression for the related Green's function. We may start with the state already containing a particle, and then we may annihilate the particle at $\mathbf{r}'$ at t' and evaluate the probability amplitude that the lack of the particle is "felt" at $\mathbf{r}$ at time t. The scalar product in (76) would be, for this case,

$$\langle \Phi_H|\Psi^+(\mathbf{r},t)\Psi(\mathbf{r}',t')|\Phi_H\rangle, \tag{77}$$

or, following the most popular convention, its complex conjugate

$$\langle \Phi_H|\Psi^+(\mathbf{r}',t')\Psi(\mathbf{r},t)|\Phi_H\rangle \,. \tag{78}$$

This expression, however, is not proportional to the evolution operator. The product $\Psi^+\Psi$ in (78) must be replaced with the commutator (for bosons) or anticommutator (for fermions) of the two field operators. In order to understand in physical terms why such a substitution is necessary, let us recall that the expectation value of the product $\Psi^+(\mathbf{r},t)\Psi(\mathbf{r},t)$ at equal positions and times gives the intensity of the field (i.e., the average number of particles at $\mathbf{r}$) for the state under consideration. The same product at different arguments gives, in the same way, the correlation between the two amplitudes at different positions and times. If we look for the dynamical correlation, that is, the propagator, without the information on the field intensity, we must subtract the product in reverse order. We may compare this result with the more familiar relations for harmonic oscillator creation, annihilation, and number operators:

$$a^+a = N \quad , \quad aa^+ = N+1, \quad \text{so that} \quad [a,a^+] = 1 \,. \tag{79}$$

From the above considerations we obtain the Green's functions:

$$G^{(r)}(\mathbf{r},t;\mathbf{r}',t') = \frac{\theta(t-t')}{i\hbar}\langle \Phi_H|[\Psi(\mathbf{r},t),\Psi^+(\mathbf{r}',t')]_\pm|\Phi_H\rangle, \tag{80}$$

$$G^{(a)}(\mathbf{r},t;\mathbf{r}',t') = -\frac{\theta(t'-t)}{i\hbar}\langle \Phi_H|[\Psi(\mathbf{r},t),\Psi^+(\mathbf{r}',t')]_\pm|\Phi_H\rangle. \tag{81}$$

In a many-body system, the Green's operators defined in (72) and (74) depend upon all the coordinates of the system, and therefore they are not of practical use. As for the density-matrix operator, however, when single-particle properties of the system are investigated, one can use single-particle Green's functions defined by the same (80) and (81), where now the vector $|\Phi_H\rangle$ describes the many-particle sysyem. They contain a reduction of the many degrees of freedom and include the effect, on each particle, of the interaction with all the other ones in the system. In the case of a system with only one particle, the operators in the reverse order do not contribute in (76) so that this equation is consistent with the general definition in (80).

As already seen above for the single-particle system, the average values of the single products of field operators in the commutators of (80) and (81) carry information about the state of the many-particle system. They are also defined as Green's functions, or, more properly, as the correlation functions $G^>$ and $G^<$:

$$G^>(\mathbf{r},t;\mathbf{r}',t') = \frac{1}{i\hbar}\langle \Phi_H|\Psi(\mathbf{r},t),\Psi^+(\mathbf{r}',t')|\Phi_H\rangle, \tag{82}$$

$$G^<(\mathbf{r},t;\mathbf{r}',t') = \mp\frac{1}{i\hbar}\langle \Phi_H|\Psi^+(\mathbf{r}',t'),\Psi(\mathbf{r},t)|\Phi_H\rangle. \tag{83}$$

If the system under consideration is only partially known, the Green's functions have to be defined as ensemble averages of the quantities defined above.

When writing an equation of motion for a single-particle Green's function, the commutator of the Hamiltonian with products of two field operators will be considered, and such commutators contain the product of four field operators, that is two-particle Green's functions. Thus, as we have already seen for the classical case and for the density matrix, a set of hierarchical equations of motion is obtained (Rickayzen, 1980), where the equation for the single-particle Green's function contains also the two-particle Green's function, the equation for the two-particle Green's function contains the three-particle Green's function, and so on.

The analysis of the single-particle Green's functions can be performed, however, by means of a general technique, based on perturbation theory. The starting point of this technique is a perturbative expansion of the various single-particle Green's functions written in the interaction picture in powers of the perturbation Hamiltonian. The various terms are commonly expressed by means of Feynman diagrams. From the analysis of such diagrams it is possible to derive a set of equations, called Dyson equations, in terms of a so-called self-energy Σ that describes the effect of the perturbation on the single-particle state. These concepts will be developed elsewhere in the ASI. We simply note that the Dyson equation has been the starting point for many approaches to quantum transport problems. As an exemple we may mention the so-called *quantum Boltzmann equation* (Mahan, 1990).

From Green's functions to density matrix and Wigner function

We have introduced many concepts of quantum theory, in the previous pages: density matrix, Wigner function, Green's functions. Our purpose, now, is to show how they are all strictly related to each other.

From the definition of $G^<$ in (83) and from the discussion related to (78), we know that $G^<(\boldsymbol{r},t,\boldsymbol{r}',t')$ represents the correlation between our particle field in $\boldsymbol{r}$ at t and the same field in $\boldsymbol{r}'$ at t'. If we evaluate such a function for $t = t'$, we obtain the instantaneous correlation of our field at the various points, and this is the one-particle density matrix, as can be seen in (43):

$$G^<(\boldsymbol{r},t;\boldsymbol{r}',t) = \frac{i}{\hbar}\overline{\langle \Psi^+(\boldsymbol{r}',t'),\Psi(\boldsymbol{r},t)\rangle} = \frac{i}{\hbar}\rho(\boldsymbol{r},\boldsymbol{r}',t)\ . \tag{84}$$

In general, the above correlation depends upon the two points $\boldsymbol{r}$ and $\boldsymbol{r}'$ separately, and not only upon their difference. We can separate the dependence on the distance between the two points from the dependence on where in space these two points are located, by the change of variables

$$\boldsymbol{R} = \frac{1}{2}(\boldsymbol{r}+\boldsymbol{r}') \quad ; \qquad \boldsymbol{s} = \boldsymbol{r}-\boldsymbol{r}' \quad ; \qquad \rho(\boldsymbol{r},\boldsymbol{r}',t) = \rho'(\boldsymbol{R},\boldsymbol{s},t)\ . \tag{85}$$

The dependence of ρ' upon $\boldsymbol{R}$ tells us how the correlation changes in various regions of space, while its dependence upon $\boldsymbol{s}$ indicates how the correlation behaves aroud $\boldsymbol{R}$. In a homogeneous system we expect ρ' to be independent of $\boldsymbol{R}$, but not of $\boldsymbol{s}$. From the above it is clear why $\boldsymbol{R}$ and $\boldsymbol{s}$ are called *slow* and *fast* space variables. The local dependence upon $\boldsymbol{s}$ will tell us something about the momentum content of the field around $\boldsymbol{R}$ (not *in* $\boldsymbol{R}$, that would be against the uncertainty principle; it is interesting to note, in this respect, that $\boldsymbol{R}$ and the differentiation with respect to $\boldsymbol{s}$, instead, do commute). In fact we may consider the Fourier transform with respect to $\boldsymbol{s}$, and it is easy to see that in this way the Wigner function is obtained, as defined in (45):

$$\frac{1}{(2\pi)^3}\int d\boldsymbol{s}e^{-i\boldsymbol{k}\boldsymbol{s}}\rho'(\boldsymbol{R},\boldsymbol{s},t) = \frac{1}{(2\pi)^3}\int d\boldsymbol{s}e^{-i\boldsymbol{k}\boldsymbol{s}}\rho(\boldsymbol{R}+\boldsymbol{s}/2,\boldsymbol{R}-\boldsymbol{s}/2,t) = f_W(\boldsymbol{R},\boldsymbol{k},t). \tag{86}$$

For completeness, we may note that the same transformation to slow and fast variables is performed also with time variables, when two-time Green's functions are considered. In that case, the slow, average, T carries the macroscopic time variations, while the difference-time τ will carry the fast quantum oscillations, whose Fourier transform corresponds to the energy.

MONTE CARLO TECHNIQUES FOR TRANSPORT EQUATIONS

A mathematical digression

In this section a few mathematical techniques for MC evaluations of sums and integrals will be reviewed. They will be used for the development of the Monte Carlo approach to classical and quantum transport.

Monte Carlo evaluation of a sum. Given the sum

$$S = \sum_i a_i\ , \tag{87}$$

a possible Monte Carlo algorithm for its evaluation is the following: a set of arbitrary probabilities p_i are defined, subject to the conditions

$$p_i \geq 0 \qquad (p_i > 0 \text{ if } a_i \neq 0), \quad i = 1,2,\ldots, \quad \sum_i p_i = 1\ . \tag{88}$$

Then a term a_i is selected with probabilities p_i, and the estimator

$$s = \frac{a_i}{p_i} \tag{89}$$

is evaluated. This is a correct estimator of the sum S; in fact its expectation value is

$$\langle s\rangle = \sum_i p_i \frac{a_i}{p_i} = S\ . \tag{90}$$

The above algorithm can be easily generalized in several ways.

Generalization to a number of sum. If, instead of a single sum S, we have to evaluate a set of sums

$$S_k = \sum_i a_{ki}\ , \tag{91}$$

a set of arbitrary probabilities p_{ki} are defined, subject to the conditions

$$p_{ki} \geq 0 \quad (p_{ki} > 0 \text{ if } a_{ki} \neq 0); \qquad \sum_{ki} p_{ki} = 1\ . \tag{92}$$

Then a term a_{kj} is selected with the probabilities p_{kj} and the estimators

$$s_j = \frac{a_{kj}}{p_{kj}}\delta_{kj} \tag{93}$$

are evaluated. These are correct estimators of the sums S_j: in fact their expectetion values are

$$\langle s_j\rangle = \sum_{ki} p_{ki}\frac{a_{ki}}{p_{ki}}\delta_{kj} = \sum_i a_{ji} = S_j\ . \tag{94}$$

Let us point out that the selection of a single term of the matrix a_{kj} yields an estimate of all the sums in (91): this estimate is a_{kj}/p_{kj} for the kth sum and zero for the other sums.

Generalization to integrals Another possible generalization of the algorithm is the substitution of a discrete sum by a continuous integral. For this generalization let us substitute S in (87) by the integral

$$f = \int_{y_1}^{y_2} g(y)dy. \tag{95}$$

Then we define an arbitrary probability density $p(y)$ subject to the conditions

$$p(y)>0 \ , \ \int_{y_1}^{y_2} p(y)dy=1; \tag{96}$$

with such a probability density we generate a value y' and evaluate the estimator

$$\frac{g(y')}{p(y')} \ . \tag{97}$$

Its expectation value is

$$\left\langle \frac{g(y')}{p(y')} \right\rangle = \int_{y_1}^{y_2} dy\, p(y') \frac{g(y')}{p(y')} = \int_{y_1}^{y_2} g(y)dy = f \ . \tag{98}$$

The corresponding generalization of the algorithm in the previous section is the evaluation of a function of the variable x defined as an integral

$$f(x) = \int_{y_1}^{y_2} g(x,y)dy. \tag{99}$$

We then define an arbitrary probability density $p(x,y)$; with such a probability density we generate a pair of values x' and y'; finally we evaluate the estimator

$$\frac{g(x',y')}{p(x',y')}\delta(x-x') \ . \tag{100}$$

Its expectation value is

$$\left\langle \frac{g(x',y')}{p(x',y')}\delta(x-x') \right\rangle = \int_{y_1}^{y_2} dy \int dx'\, p(x',y') \frac{g(x',y')}{p(x',y')}\delta(x-x')$$
$$= \int dx'\, f(x')\delta(x-x') = f(x) \ . \tag{101}$$

Generalization to a function defined as an infinitive sum of multiple integrals. The final generalization of the algorithm considers the evaluation of a function defined as one series of multiple integrals:

$$f(x) = \int dy_0 g(x,y_1) + \int dy_1 \int dy_2 g(x,y_1,y_2) + \int dy_1 \int dy_2 \int dy_3 g(x,y_1,y_2,y_3) + \cdots \tag{102}$$

For such a case the algorithm will consist of a choice, with given arbitrary probability, of a particular multiple integral in the sum and of the point where the integrand function is evaluated. If the nth integral and the point $(x',y_1',y_2',\ldots,y_n')$ are chosen, then the estimator is

$$\frac{g(x',y_1',y_2',\cdots,y_n')}{p(x',y_1',y_2',\cdots,y_n')}\delta(x-x'). \tag{103}$$

where

$$p(x',y_1',y_2',\cdots,y_n') \tag{104}$$

is the probability of the realized choice.

Let us emphasize that the probabilities used in these algorithms are arbitrary. The correctness of the estimator does not depend on them. The variance of the estimator, on the other hand, does. For example, if the evaluation of the sum in (87) is performed by using as probability for the ith term its relative contribution to the sum (which is possible if all terms have the same sign):

$$p_i = \frac{a_i}{S} \ , \tag{105}$$

then each estimate would lead to the exact value of the sum:

$$\frac{a_i}{p_i} = \frac{a_i}{a_i/S} = S \ , \tag{106}$$

and we would have a vanishing variance. Such a situation cannot be realized, of course, since we do not know *a priori* the value of the sum S to be estimated. The example, however, illustrates the fact that a smart choice of the probabilities may reduce significantly the variance of the results.

MC solution of Boltzmann equation

Integral Chambers form of the Boltzmann equation. As a starting point, let us recall the explicit form of the B.E. discussed above:

$$\mathrm{L}f = \frac{\partial f}{\partial t} + v(\mathbf{k})\frac{\partial f}{\partial \mathbf{r}} + \frac{F}{\hbar}\frac{\partial f}{\partial \mathbf{k}} = \frac{V}{(2\pi)^3}\int d\mathbf{k}'\{f(\mathbf{k}')P(\mathbf{k}',\mathbf{k}) - f(\mathbf{k})P(\mathbf{k},\mathbf{k}')\}. \tag{107}$$

Now we change variables from $(\mathbf{r},\mathbf{k},t)$ to $(\mathbf{r}^*,\mathbf{k}^*,t^*)$, where $\mathbf{r}^*$ and $\mathbf{k}^*$ are those defined after (60) ($\mathbf{k}$ in place of ν does not introduce any essential difference), and $t^* = t$. The distribution function in the new variables will be

$$f^*(\mathbf{r}^*,\mathbf{k}^*,t^*) = f(\mathbf{r}(\mathbf{r}^*,\mathbf{k}^*,t^*),\ \mathbf{k}(\mathbf{r}^*,\mathbf{k}^*,t^*),\ t(\mathbf{r}^*,\mathbf{k}^*,t^*)). \tag{108}$$

By means of (63) it is straightforward to show that

$$\mathrm{L}f = \frac{\partial}{\partial t^*} f^*(\mathbf{r}^*,\mathbf{k}^*,t^*) \ . \tag{109}$$

The above result allows us to write the BE for the transformed distribution function f^* as

$$\frac{\partial}{\partial t^*} f^*(\mathbf{r}^*,\mathbf{k}^*,t^*) = \frac{V}{(2\pi)^3}\int d\mathbf{k}'\{f^*(\mathbf{r}^*,\mathbf{k}^*,t^*)P(\mathbf{k}',\mathbf{k}) - f^*(\mathbf{r}^*,\mathbf{k}^*,t^*)P(\mathbf{k},\mathbf{k}')\}. \tag{110}$$

Here, now, k is a function of t since $k = k(r^*,k^*,t)$ and in the following it will be simply indicated by $k(t)$. By introducing the total scattering probability, defined as the integral of the scattering probability P over all the final states k'

$$\gamma(k(t)) = \frac{V}{(2\pi)^3}\int P(k(t),k')dk' \quad , \tag{111}$$

the transport equation (107) may be written as

$$\frac{\partial}{\partial t} f^*(r^*,k^*,t^*) = -\gamma(k(t))\, f^*(r^*,k^*,t^*) + \frac{V}{(2\pi)^3}\int dk' f^*(r^*,k^*,t^*)P(k',k). \tag{112}$$

This can be written as

$$e^{-\int_{t_0}^{t}\gamma(k(t'))dt'}\frac{\partial}{\partial t}\left\{ f^*(r^*,k^*,t^*)e^{\int_{t_0}^{t}\gamma(k(t'))dt'}\right\} = \frac{V}{(2\pi)^3}\int dk' f^*(r^*,k^*,t^*)P(k',k) \,, \tag{113}$$

or, after a formal time integration,

$$\begin{aligned} f^*(r^*,k^*,t) &= f^*(r^*,k^*,t_0)e^{-\int_{t_0}^{t}\gamma(k(t'))dt'} \\ &+\int_{t_0}^{t} dt' e^{-\int_{t'}^{t}\gamma(k(t''))dt''} \times \frac{V}{(2\pi)^3}\int dk' f^*(r^*,k^*,t')P(k',k(t')). \end{aligned} \tag{114}$$

We may finally go back to the original variables and obtain

$$\begin{aligned} f(r,k,t) &= f(r(t_0),k(t_0),t_0)e^{-\int_{t_0}^{t}\gamma(k(t'))dt'} \\ &+\int_{t_0}^{t} dt' e^{-\int_{t'}^{t}\gamma(k(t''))dt''} \times \frac{V}{(2\pi)^3}\int dk' f(r(t'),k',t')P(k',k(t')) \end{aligned} \quad , \tag{115}$$

where now r, k and t are again the independent variables of the equation and $r(t')$ and $k(t')$ are the ballistic position and momentum of a particle that is at r and k at time t.

This is Chambers integral version of the transport equation. It has a very straightforward physical interpretation: the distribution function $f(r,k,t)$ is given by two contributions: the first one is given by the particles that at $t=t_0$ are already in the right trajectory and are not scattered away from it before time t; the second contribution is given by the particles that are put in the right trajectory in r' at any time t' between the initial time and t with the right momentum k' and are not scattered away before time t.

Series expansion of Chambers equation. In order to simplify the notation, let us introduce the damping factor

$$S(t,t') = e^{-\int_{t'}^{t}\gamma(k(\tau))d\tau} \quad . \tag{116}$$

In terms of this factor S and of the in-scattering operator defined by

$$P_i(t)f(r,k,t) = \frac{V}{(2\pi)^3}\int dk' f(r(t'),k',t)P(k',k(t)) \quad , \tag{117}$$

the Chambers equation (115) can be simply rewritten as

$$f(r,k,t) = S(t,t_0)f(r(t_0),k(t_0),t_0) + \int_{t_0}^{t} dt' S(t,t')P_i(t')\, f(r(t'),k',t') \quad . \tag{118}$$

If we now substitute iteratively the above integral equation into itself, we obtain a series expansion for f. With a simplified notation, the expansion is written as

$$\begin{aligned} f(t) &= S(t,t_0)f(t_0) + \int_{t_0}^{t} dt_1 S(t,t_1)P_i(t_1)S(t_1,t_0)\, f(t_0) \\ &+\int_{t_0}^{t} dt_1 \int_{t_0}^{t_1} dt_2\, S(t,t_1)P_i(t_1)S(t_1,t_2)P_i(t_2)S(t_2,t_0)\, f(t_0)+\ldots \end{aligned} \quad . \tag{119}$$

The explicit form of the nth order term is

$$\begin{aligned} \Delta f^{(n)}(r,k,t) &= \int_{t_0}^{t} dt_1 \int_{t_0}^{t_1} dt_2 \ldots \int_{t_0}^{t_{n-1}} dt_n \left(\frac{V}{(2\pi)^3}\right)^n \int dk_1 \int dk_2 \ldots \int dk_n \\ &\times e^{-\int_{t_1}^{t}\gamma(k(\tau))d\tau} P(k_1(t_1),k(t_1)) e^{-\int_{t_2}^{t_1}\gamma(k_1(\tau))d\tau} P(k_2(t_2),k_1(t_2))\ldots \\ &\times e^{-\int_{t_n}^{t_{n-1}}\gamma(k_{n-1}(\tau))d\tau} P(k_0(t_n),k_{n-1}(t_n)) e^{-\int_{t_0}^{t_n}\gamma(k_0(\tau))d\tau} f(r_0,k_0,t_0) \end{aligned} \quad , \tag{120}$$

where $k_0=k_n$. The above expression can be interpreted as a series of scattering events in phase space at the various times $t_n,t_{n-1},\ldots,t_1$. It involves only in-scattering processes, and, for each pair of adjacent processes at times t_i and t_j the damping factor $S(t_i,t_j)$ is inserted. This is the probability of a free flight with duration t_i - t_j (i.e. the probability that a particle travels without scattering from t_i to t_j. If we had started with the BE instead of the Chambers equation, out-scattering events would have also contributed, and the damping factor would have appeared as the result of the infinite summation of all possible sequences of out-scattering events between two successive in-scatterings (Rossi *et al.*, 1992).

In conclusion, we can regard each term of the present expansion as a particular particle trajectory in space from the initial time to the current time t.

MC solution of Boltzmann equation. The Monte Carlo (MC) solution of the Boltzmann transport equation for electron in semiconductors is well known as a direct simulation of the semiclassical motion of the electrons inside the crystal. Here we shall present this approach in more general terms, suitable to a generalization to quantum transport. The present method is derived by applying the general MC algorithm for the evaluation of an infinite sum discussed previously. For the case of various scattering mechanisms, the series expansion for f can be written as

$$f(t) = S(t,t_0)f(t_0) + \int_{t_0}^{t} dt_1 S(t,t_1) \sum_{i_1} P^{(i_1)}(t_1) S(t_1,t_0) f(t_0)$$
$$+ \int_{t_0}^{t} dt_1 \int_{t_0}^{t_1} dt_2 S(t,t_1) \sum_{i_1} P^{(i_1)}(t_1) S(t_1,t_2) \sum_{i_2} P^{(i_2)}(t_2) S(t_2,t_0) f(t_0) + \ldots \quad (121)$$
$$= f^{(0)}(t) + \Delta f^{(1)}(t) + \Delta f^{(2)}(t) + \ldots$$

where the explicit sums over the various scattering mechanisms with scattering rates $P^{(i)}$ have been introduced. Then, the general MC algorithm described above provides an estimate of the distribution function f by means of random selections of the various terms of the above expansion. Let us recall that also the operators $P^{(i)}$ contain sums over the final scattering states and these will also be estimated by MC selections. Thus, each selected term will correspond to a possible electron trajectory, and the selections can be made with arbitrary probabilities. For example, the following algorithm can be used:

(i) The initial electron variables r_0 and k_0 are chosen with arbitrary probability distribution $p_0(r_0,k_0)$.

(ii) The number n of scattering events (i.e. the perturbative order in the expansion) and the times t_1, t_2, ..., t_n at which the integrand function has to be evaluated (i.e. the times of the scattering events) are randomly selected with arbitrary probability distribution $p_t(n, t_1, t_2, ..., t_n)$

(iii) For each of the n scattering events, we perform a random selection of the scattering mechanism s and of the corresponding final state k' with arbitrary probabilities $p_s(s,k')$.
The above choices correspond to the estimator

$$\left\langle \frac{e^{-\int_{t_0}^{t}\gamma(\mathbf{k}(\tau))d\tau} P^{(s_1)}(\mathbf{k}_0(t_1),\mathbf{k}(t_1)) \ldots P^{(s_n)}(\mathbf{k}_0(t_n),\mathbf{k}_{n-1}(t_n)) e^{-\int_{t_0}^{t_n}\gamma(\mathbf{k}_0(\tau))d\tau} f(\mathbf{r}_0,\mathbf{k}_0,t_0)}{p_t(n,t_1,\ldots,t_n)\, p_s(s_1,\mathbf{k}(t_1)) \ldots p_s(s_n,\mathbf{k}_{n-1}(t_n))\, p_0(\mathbf{r}_0,\mathbf{k}_0)} \right\rangle . \quad (122)$$

As for the case of traditional MC simulation, the self scattering technique can be employed in order to simplify the form of the above estimator. Furthermore, as a particular choice, we can employ the standard free-flight generation with arbitrary lifetime γ_0^{-1}. With these choices, denoting Γ_0 the total scattering rate, including self scattering, the estimator in (122) reduce to

$$\left\langle \frac{e^{-\Gamma_0(t-t_0)} P^{(s_1)}(\mathbf{k}_0(t_1),\mathbf{k}(t_1)) \ldots P^{(s_n)}(\mathbf{k}_0(t_n),\mathbf{k}_{n-1}(t_n)) f(\mathbf{r}_0,\mathbf{k}_0,t_0)}{e^{-\gamma_0(t-t_0)}\gamma_0^{\,n}\, p_s(s_1,\mathbf{k}(t_1)) \ldots p_s(s_n,\mathbf{k}_{n-1}(t_n))\, p_0(\mathbf{r}_0,\mathbf{k}_0)} \right\rangle . \quad (123)$$

From the point of view of electron-transport simulation, the resulting algorithm proceeds as follows:

(I) At time t_0 the initial electron state (r_0,k_0) is selected. The corresponding estimator is initialized with the value $f(r_0,k_0,t_0)/p_0(r_0,k_0)$.

(ii) A random free flight is generated according to the arbitrary total probability γ_0, as $t_r = -\gamma_0^{-1} \log(r)$ leading to the value $t_n = t_0 + t_r$ for the scattering event; the current estimator is multiplied by the factor $\exp(-(\Gamma_0 - \gamma_0))$.

(iii) The mechanism s_n responsible for the scattering event and the final state $k_{n-1}(t_n)$ of the scattering are randomly selected according to the arbitrary probability $p_s(s,k')$; the current estimator is multiplied by the ratio

$$P^{(s_n)}(\mathbf{k}_0(t_n),\mathbf{k}_{n-1}(t_n)) / \gamma_0 p_s(s_n,\mathbf{k}_{n-1}(t_n)) . \quad (124)$$

(iv) The electron simulation repeats points (ii) and (iii) until the final time is reached. At this point the resulting estimator, related to the final phase-space coordinates (r_f,k_f), is added to the corresponding counter for the evaluation of f.

If we employ as arbitrary total scattering rate γ_0 the natural total scattering rate Γ_0, as arbitrary probabilities for the various scattering mechanisms p_s the natural relative probabilities $P^{(s)}/\Gamma_0$ and, in addition, we use as arbitrary probability $p_0(r_0,k_0)$ the initial distribution f_0, the estimator in (123) reduces to unity, the evaluation of f is obtained by simply counting the number of electrons reaching a given final state with "natural" probabilities, and the traditional Ensemble MC is recovered.

On the other hand, the freedom related to the arbitrariness of the probabilities of each choice can be used to "guide" the electrons toward regions of phase space of particular interest. More electrons (than in real physics) will sample this region with an average weight less than unity leading to an unbiased estimator with reduced variance.

MC solution of the Liouville-vonNeumann equation (Brunetti et al., 1989)

As already indicated, the general MC approach applied above to the BE, can be used also for quantum-transport equations. We shall first consider the Liouville-von Neumann equation in its general form (31), with electron and phonon coordinates included. However, since we are not going to include electron electron interaction, we shall consider a system formed by a single electron (or many independent electrons) and the phonon field, with an applied, constant and uniform electric field E. The total hamiltonian is

$$\mathbf{H} = \mathbf{H}_e + \mathbf{H}_p + \mathbf{H}_E + \mathbf{H}_i \quad (125)$$

where $\mathbf{H}_e$ is the free electron Hamiltonian, $\mathbf{H}_p$ the free phonon Hamiltonian, $\mathbf{H}_E$ the contribution of the applied field, and

$$\mathbf{H}_i = \sum_q i\hbar F(q)\left\{a_q e^{iqr} - a_q^{+} e^{-iqr}\right\} \quad (126)$$

is the electron-phonon interaction given in terms of phonon creation and anihilation operators.

Integral form of Liouville-vonNeumann equation. Let us consider the set of basis vectors

$$|x,t\rangle = |\mathbf{k}_0,t\rangle |\{n_q\},t\rangle , \quad (127)$$

where

$$\langle \mathbf{r}|\mathbf{k}_0,t\rangle = \frac{1}{\sqrt{V}} e^{i\mathbf{k}(t)\cdot\mathbf{r}}\, e^{-i\int_{t_0}^{t} d\tau\, \omega(\tau)} \quad (128)$$

are accelerated plane waves, being

$$\mathbf{k}(t) = \mathbf{k}_0 - e\frac{\mathbf{E}}{\hbar}(t - t_i) \quad ; \qquad \omega(t) = \frac{\hbar \mathbf{k}(t)^2}{2m} , \quad (129)$$

and $|\{n_q\},t\rangle$ are the phonon number eigenstates with their time dependence. In terms of these basis vectors, (31) assumes the same form as if the interaction picture were used with $\mathbf{H}_e + \mathbf{H}_E + \mathbf{H}_p$ as unperturbed Hamiltonian:

$$\frac{\partial}{\partial t}\rho(x,x',t) = \sum_{x''}\{H'(x,x'',t)\rho(x'',x',t) - \rho(x,x'',t)H'(x'',x',t)\}\ , \tag{130}$$

where

$$H'(x,x',t) = \frac{1}{i\hbar}\langle x,t|\mathbf{H}_i|x',t\rangle = -H'^{*}(x',x,t)\ . \tag{131}$$

Equation (130) is the integro-differential equation for the density matrix. We can describe the equation by saying that the matrix elements of the interaction Hamiltonian act twice on the density matrix, one time per argument; on the first argument it acts as it is, on the second argument it acts as complex conjugate. By formal integration, we obtain the integral equation

$$\rho(x,x',t) = \rho(x,x',t_i) + \int_{t_i}^{t} dt' \sum_{x''}\{H'(x,x'',t')\rho(x'',x',t') - \rho(x,x'',t')H'(x'',x',t')\}\ , \tag{132}$$

where the matrix elements are different from zero only if one mode q has occupation number changed by one unity and the electron wavevector is changed by the corresponding q. This is the integral equation to be solved with the iterative expansion and the MC technique.

Series expansion of Liouville-vonNeumann equation. The series expansion of Eq. (132) is obtained, as usual, by substituting the equation into itself. We assume that the density matrix at the initial time is factorized into an electron part and a phonon part, and that they correspond to diagonal equilibrium distributions for the corresponding free Hamiltonians. Furthermore, we are interested in the diagonal elements of the density matrix since we shall perform the trace over the phonon coordinates and we look for diagonal elements of the electron density matrix. As a consequence:

(I) Only the even order terms of the series expansion give contributions.

(ii) A mode q which has been absorbed (or emmited) by one argument must be absorbed (or emited) by the other argument, or reemited (or reabsorbed) by the same argument. Each couple of paired matrix elements (vertices) will be called a process and corresponds to a scattering event.

MC solution of the Liouville-vonNeumann equation. The numerical procedure starts with random selections, with suitable arbitrary probabilities, of the order of the perturbative correction, the sequence of emission and absorbtion processes, the times of the processes in the sequence and the modes q of the involved phonons.

Then, starting from the value $k(t)$ at which ρ is to be evaluated, both arguments of the density matrix are translated backwards in time until the time t_1 of the latest vertex. At this point the contribution of the corresponding matrix element H' is evaluated and the current value of k is changed accordingly. This step is repeated until all involved processes are considered and the initial values of the arguments of ρ are reached.

Finally, according to the general theory, the estimator

$$\frac{H'(t_1)H'(t_2)\ldots H'(t_N)}{P}\rho(x_i,x_i,t_i) \tag{133}$$

is evaluated, where P is the product of the single probabilities used above. This estimator will contain a factor n_q or n_q+1 for each process of absorption or emission, respectively. We also assume that each mode q appears only once in a given perturbation term (no hot phonons), so that the corrections in ρ are linear with respect to n_q, and the trace over the phonons results in replacing n_q with its equilibrium value. In this way we obtain the density matrix trough a random generation of quantum processes in the same way as tradition ensemble MC obtains the distribution function trough a random generation of electron trajectories.

MC solution of the Wigner equation (Rossi *et al.*, 1994)

The Wigner equation is particularly useful when quantum transport is studied in space-dependent problems. A space-dependent field may be applied, corresponding to a potential $V(r)$, and boundary conditions can be specified (Frensley, 1990). We shall not repeat again all the description for the MC solution of the Wigner equation. It follows the same procedure as the two previous cases. Equation (53) is first transformed by means of path variables so that the l.h.s. contains only the time derivative. The r.h.s. contains the contributions of the transfer function $W(r,k)$ defined in (50), as well as phonon matrix elements as shown in (53). Then a formal integration yields an integral equation that can be developped in a series expansion with iterative substitution of the equation into itself. As before, we are interested in obtaining diagonal elements with respect to the phonon states. The series will contain three types of terms. Terms of the first type contain contributions of the transfer function $W(r,k)$ only and are diagonal with respect to phonons. They describe the coherent electron dynamics in presence of the applied potential. The terms of second type contain phonon processes as described above for the Liouville equation, and the terms of the last type contain mixed interactions. These last terms describe the combined effects of the applied field and phonon scattering (intracollisional field effect). The convergence of the series has been discussed by Nedjalkov *et al.* (1994). The MC procedure follows the same lines as before.

CLASSICAL LIMIT OF THE QUANTUM TRANSPORT EQUATIONS

The derivation of the semiclassical limit from quantum transport equation can be approached in many different ways. Here we would like to sketch a method strictly related to the series expantion of the Wigner equation developped in the previous sections.

We already stated in connection with quantum dynamics that the field variations must be slow compared with electron wavelegth in order to be able to use classical dynamics. In such a case the applied field can be put on the l.h.s. of the Wigner equation (51) so that on the r.h.s. only scattering agents are left. In the weak coupling limit only separate processes have to be considered and the time integration of one of their two vertices can be considered as extended to infinity so that the completed collision limit is reached. The delta function of energy conservation is thus recovered together with the Fermi golden rule for the scattering. In this way the series that results coincides with what is obtained from Boltzmann equation.

WAVEGUIDES AND THE SCHRÖDINGER EQUATION

At this point, we now want to turn to the steady-state analysis of propagation through quantum waveguide structures. Although there are different formulations of

quantum, nearly all approaches which lead to modeling of semiconductor devices derive from the Schrödinger equation

$$i\hbar\frac{\partial\Psi}{\partial t}=-\frac{\hbar^2}{2m}\nabla^2\Psi+V(x,y,z)\Psi \ , \tag{134}$$

in three dimensions. As discussed in the previous sections, modeling can proceed from either this equation, or a variety of forms that are derived from this equation. It must be noted here that if we want to use this equation itself in modeling, we must do so for a sufficiently small structure, since (134) does not include any dissipative terms. For this to be accurate, one must consider a great variety of characteristic lengths.

Lengths of interest always begin by a consideration of the scattering that occurs within the "device" of interest. There are two basic types of scattering that are important: (1) elastic scattering that induces a mean time between scattering τ, although there are times when modifications of this are introduced to produced a transport relaxation time τ_e, and (2) the inelastic scattering time τ_{in}, which describes the energy, or phase, relaxation process in the device. These times are then coupled to the lengths of interest either through the average velocity of the carriers, either the Fermi velocity v_F or the thermal velocity v_T, or through the diffusion coefficient $D = v_F^2\tau\,/\,d$, where d is the dimensionality of the system, and when the dominant average velocity is the Fermi velocity. At relatively low temperatures, where one predominantly studies *mesoscopic* systems, the carriers are degenerate, and the Fermi velocity is the relevant average velocity. At higher temperatures, the situation is more complicated, but here we will assume that the former is the case.

The *elastic mean free path* l is defined as the average length traveled in the time τ, which becomes $l = v_F\tau$. Similarly, we can define the *inelastic mean free time* $l_{in} = v_F\tau_{in}$. In mesoscopic systems, where the transport is predominantly diffusive (large degree of disorder), the *phase coherence length* is defined as $l_\phi = (D\tau_{in})^{1/2}$ (Thouless, 1977). Many people consider these latter two times to be equal, but this is not the case! From the definitions above, for example, we can express the ratio as

$$\frac{l_{in}}{l_\phi}=\frac{v_F\tau_{in}}{\sqrt{D\tau_{in}}}=\sqrt{d\frac{\tau_{in}}{\tau}} \ , \tag{135}$$

so that in general $l_{in} > l_\phi$. Typical values of these lengths are given in the chapter by Rössler, for the GaAs system. Generally, the inelastic mean free path is the important length for quasi-ballistic systems in which there is little scattering, or $L \le l_{in}$. (Here, L is the characteristic length of the quantum system.) On the other hand, the phase coherence length is the critical length for systems in which there is significant scattering and the transport is largely diffusive, $L > l_\phi, l$. We are interested in the former situation, as we want to examine the transport of *waves* through the quantum system.

Electron Waveguides

The basic concepts of transport in mesoscopic systems, in which there are few scatterers, can be traced to Landauer (1957). While the onset of significant scattering will suppress many of the quantum effects of interest, and can cause localization in the device (we return to this below), we will assume that this is not the major point of interest and neglect the scattering until later sections. The treatment of transport with the Schrödinger equation that we follow is largely based upon the assumption of transport of the particles as simple waves in quantum waveguides, with occasional scatterers are imbedded (Büttiker, 1988a, 1988b, 1989). In this, we make an analogy between electron waves and microwaves that both propagate in the appropriate waveguide.

The creation of an electron waveguide is usually carried out in a modulation-dope AlGaAs/GaAs heterostructure, in which a quasi-two-dimensional electron gas is created on the GaAs side of the hetero-interface. The waveguide can be fabricated then by either using lateral surface gates, or by physically defining the structure by reactive-ion etching (see the chapter by Reed in this volume). The direction normal to the surface is taken to be the z direction, and motion in that direction is constrained by the interface, so that we take carriers only in the lowest 2D subband of this structure. The waveguide is confined by potentials in the y direction, so that free propagation is assumed only in the x direction. Schrödinger's equation can now be written (in the time-independent form) as

$$-\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)\Psi(x,y)+V(x,y)\Psi(x,y)=E\Psi(x,y) \ , \tag{136}$$

with

$$V(x,y) = V_c(y) + V_{appl}(x,y) \ , \tag{137}$$

and the first term on the right-hand side is the confinement potential defining the lateral extent of the waveguide while the last term is any applied potential describing bias or impurities, etc. The general solution of the wave function in any small region, or in any small incremental length δL, over which the confinement potential (and the applied potential) is uniform is given by

$$\Psi(x,y)=\sum_n \phi_n(x)\chi_n(y) \ , \tag{138}$$

where, in general for hard wall boundaries in the y direction,

$$\chi_n(y)=\sqrt{\frac{2}{W}}\sin\left(\frac{n\pi y}{W}\right). \tag{139}$$

Other confinement schemes assume a quadratic variation, so that the wavefunctions χ_n are harmonic oscillator wave functions. Actually, numerical solutions suggest that for gate defined structures, the quadratic variation near the boundaries is a very good approximation (Laux *et al.*, 1988; Kumar *et al.*, 1989).

The longitudinal modes are described, in general, by a combination of *forward* and *backward* waves, as

$$\phi_n(x)=a_n e^{ik_n x}+b_n e^{-ik_n x} \ , \tag{140}$$

where k_n is the propagation constant of the n-th mode. If k_n is real, the waves are propagating waves, while if k_n is imaginary, $k_n = i\gamma_n$, the waves are damped (evanescent) waves. *It is very important to note that proper inclusion of the evanescent modes is very important in studying waveguide discontinuities by this method, just as it is for microwave waveguides.* At the interface between two regions, in each of which the mode properties are uniform, the total wave function and its derivative are matched across the interface. If local potentials are present at the interface, then the derivative is discontinuous by this amount as will be seen.

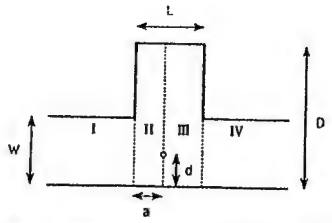


Figure 1. Schematic of a T-shaped electron waveguide with a single impurity placed in the stub region. Hard-wall confinement is assumed.

Impurity in a Resonator

As a first example of the use of the mode matching technique, we consider a waveguide stub tuner, with a single impurity located in the resonance region. The conductance of a mesoscopic system has been recognized for some time to be quite sensitive to the position of discrete, individual impurities (Ralls *et al.*, 1984). Here, we study this for a simple configuration. Consider Fig. 1. The waveguide composes regions I and IV of the figure, while the stub tuner composes regions II and III. The latter are separated by a single impurity described by the δ-function potential

$$V_{appl}(x,y) = V_{imp}(x,y) = \gamma\delta(x)\delta(y-d) \ , \tag{141}$$

where $x = 0$ is taken at the center of the stub, and γ is the amplitude (strength) of the scattering potential. The various dimensions and notation are described in the figure. The wave functions in regions II and III can be taken to be combinations of forward and backward (rightward and leftward propagating) modes as (Takagaki and Ferry, 1992a)

$$\Psi_{II} = \sum_{j=1} \sqrt{\frac{2}{D}} \sin\left(\frac{j\pi y}{D}\right)\left(A_j e^{iq_j x} + B_j e^{-iq_j x}\right), \quad -a < x < 0 \ , \tag{142}$$

and

$$\Psi_{III} = \sum_{j'=1} \sqrt{\frac{2}{D}} \sin\left(\frac{j'\pi y}{D}\right)\left(C_{j'} e^{iq_{j'} x} + F_{j'} e^{-iq_{j'} x}\right), \quad 0 < x < -a \ . \tag{143}$$

In the absence of any scatterer, $j = j'$. However, when there is a scatterer present, each mode in region II couples with all other modes in region III, and vice versa. Using the orthogonality of the sinusoidal wave functions in the transverse direction, the coupling of the wave functions at $x = 0$ leads to a modified Schrödinger equation for the x-components of the wave functions as:

$$\frac{d^2\phi_j(x)}{dx^2} + q_j^2\phi_j(x) = \sum_m \gamma_{jm}\phi_m(x) \ , \tag{144}$$

with

$$\gamma_{jm}(x) = \frac{2m}{\hbar^2}\int dy \chi_j(y) V_{imp}(x,y)\chi_m(y) \ . \tag{145}$$

In the above equations, q_j is real for propagating modes and imaginary for evanescent modes, as discussed above, with the magnitude being given by

$$\arg(q_j) = \sqrt{\left|k_F^2 - (j\pi / D)^2\right|} \ . \tag{146}$$

The modified Schrödinger equation (144) defines the continuity needed to match the solutions at $x = 0$. Integrating (144) once gives the boundary condition for the derivative of the wave function (Bagwell, 1990)

$$\left.\frac{d\phi_j(x)}{dx}\right|_{x=0^+} - \left.\frac{d\phi_j(x)}{dx}\right|_{x=0^-} = \sum_m \Gamma_{jm}\phi_m(0) \ , \tag{147}$$

where

$$\Gamma_{jm} = \frac{4m\gamma}{\hbar^2 D}\sin\left(\frac{j\pi d}{D}\right)\sin\left(\frac{m\pi d}{D}\right) \tag{148}$$

for the δ-function impurity potential. It is clear from this last equation that mode mixing depends crucially upon the position of the impurity in the waveguide. Little mixing occurs if the impurity is near one side, but maximal mixing occurs when the maximum of the modes occurs at the site of the impurity. Substituting the assumed form of the wave functions leads to the condition

$$iq_j(C_j - F_j) - iq_j(A_j - B_j) = \sum_m \Gamma_{jm}(A_m + B_m) \ . \tag{149}$$

The corresponding condition on the continuity of the wave function is represented by

$$A_j + B_j = C_j + D_j \ . \tag{150}$$

Using the above equations, and their counterparts at the points $x = \pm a$, we have solved for the transmitted and reflected wave amplitudes. Sols *et al.* (1989a, 1989b) have pointed out that it is desirable, in order to use the stub tuner action as a transistor, to have $L/W \sim 1$ and a single-mode regime. In Fig. 2, we show results of the dependence on the position of the impurity for $L/W = 1$ and $D/W = 1.6$. In part (a), the impurity is moved along the x axis with $d/W = 1.13$. The dotted, dashed, and solid curves are for $a/W = 0.26$, 0.43, and 0, respectively. In part (b), the impurity is moved along the y axis with $a/W = 0.57$. The solid, dotted, dashed, dash-dotted curves are for $d/W = 0$, 0.32. 0.68. and 1.13, respectively. In each case the horizontal axis is essentially the number of modes propagating. The reduced strength of the impurity is taken to be $U = 4m\gamma W / \hbar^2 D$.

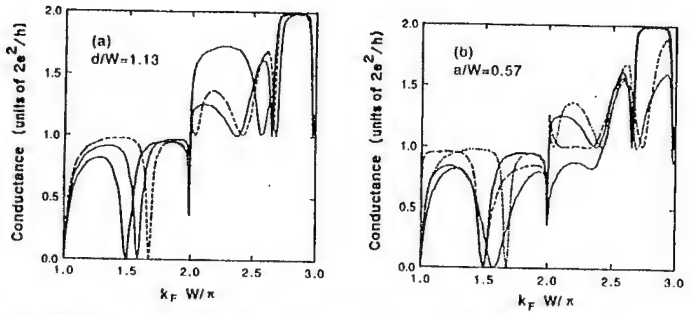


Figure 2. Two-terminal conductance as a function of $k_F W/\pi$ for various positions of an impurity with the strength U=50 for motion along the (a) x-axis and (b) y-axis. The curves are discussed in the text.

Tunneling Spectroscopy of Point Contacts

Electron transport in the quasi-ballistic regime has attracted a great deal of attention in recent years. Of particular interest is the quantum point contact in which two depletion gates are biased, so that a very short quantum wire is produced between the two gates. If the wire is sufficiently short (to be discussed by several authors in this volume), the conductance of this *point contact* is defined to be integer multiples of $2e^2/h$, as the width is varied. In the absence of scattering, each occupied subband carries the same amount of current since the group velocity and the density of states yield a product of unity in a quasi-one-dimensional system. Hence, the integer multiplier is the number of occupied subbands in the short wire that forms the point contact.

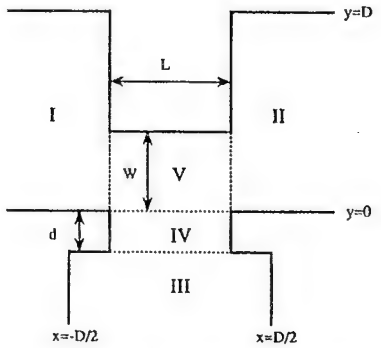


Figure 3. Schematic diagram of the point contact with a side probe for tunneling spectroscopy. A potential barrier of height U is placed in region IV.

A recent experiment on tunneling spectroscopy of a waveguide has revealed an oscillation in the current leaking out of the waveguide through a side-wall (Eugster and del Alamo, 1991). In this experiment, the waveguide was defined by a long quantum point contacts, and the authors suggested that the oscillation was due to modulation of the tunneling current as the conductance steps occured with varying gate bias on the confining gates. In the present example, we will calculate the waveguide properties of a model structure inspired by this experiment. We find that the tunneling spectroscopy is greatly affected by the actual shape of the barrier potential, and that an oscillation exists that is due to longitudinal resonant states in the constriction rather than directly to the density of states itself (Takagaki and Ferry, 1992b). The model is illustrated in Fig. 3, and consists of five regions, which are labeled I, II, ..., V. A narrow constriction (region V) forms the quantum wire under investigation, and this is coupled through a tunneling barrier (region IV) to a side lead. The dimensions are shown in the figure. As above, hard wall confinement is assumed for the side walls of all regions. The tunneling barrier is taken into account by placing a potential step with height U in this region. The tunneling spectrum crucially depends upon the height of this potential. However, the basic features of the tunneling current into the sidearm are given well by this square potential barrier.

We consider an electron with energy $E_F=\hbar^2k^2/2m$ incident through mode n in lead I, on the left of the figure. In the uniform waveguide sections, the wave functions can be given by

$$\Psi_{I,n}(x,y)=\sum_m(\delta_{mn}e^{ik_mx}+R_{mn}e^{-ik_mx})\sin\left(\frac{m\pi y}{D}\right), \tag{151}$$

$$\Psi_{II,n}(x,y)=\sum_m T_{mn}e^{ik_mx}\sin\left(\frac{m\pi y}{D}\right), \tag{152}$$

$$\Psi_{III,n}(x,y)=\sum_m S_{mn}e^{-ik_my}\sin\left(\frac{m\pi}{D}(x+\frac{D}{2})\right), \tag{153}$$

$$\Psi_{IV,n}(x,y)=\sum_m(A_{mn}e^{-p_my}+B_{mn}e^{p_my})\sin\left[\frac{m\pi}{L}\left(x+\frac{L}{2}\right)\right], \tag{154}$$

where the longitudinal momentum wave vectors are given by

$$k_m=\sqrt{k_F^2-\left(\frac{m\pi}{D}\right)^2}\ ,\quad p_m=\sqrt{\left(\frac{m\pi}{L}\right)^2+\frac{2m^*U}{\hbar^2}-k_F^2}\ . \tag{155}$$

In the above equations, we have chosen U such that the longitudinal momenta of the tunneling electrons take imaginary values (the Fermi level lies below the peak of the barrier in region IV). Finally, in region V, the wave function is given as an expansion of standing waves of the form

$$\Psi_{V,n}=\sum_m(C_{mn}e^{iq_mx}+F_{mn}e^{-iq_mx})\sin\left(\frac{m\pi y}{W}\right)$$
$$+\sum_m E_{mn}\sin[r_m(y-W)]\sin\left[\frac{m\pi}{L}\left(x+\frac{L}{2}\right)\right], \tag{156}$$

where

$$q_m=\sqrt{k_F^2-\left(\frac{m\pi}{W}\right)^2}\ ,\quad r_m=\sqrt{k_F^2-\left(\frac{m\pi}{L}\right)^2}\ . \tag{157}$$

The wave functions in each region are matched across each of the interfacial boundaries. The transmission probability through the point contact (from region I to region II) T_p and the tunneling probability (from region I to region III) T_t, respectively, are obtained as

$$T_p = \sum_{m,n}(k_m / k_n)|T_{mn}|^2 \quad , \quad T_t = \sum_{m,n}(k_m / k_n)|S_{mn}|^2 \quad . \tag{158}$$

The sums run only over the occupied subbands, as the other subbands (whether propagating or evanescent) do not contribute to the current. If the voltage is applied to lead I, while leads II and III are held at zero potential, the current can then be calculated from $I_p=(e^2/h)T_pV_{appl}$, and similarly for the tunneling current.

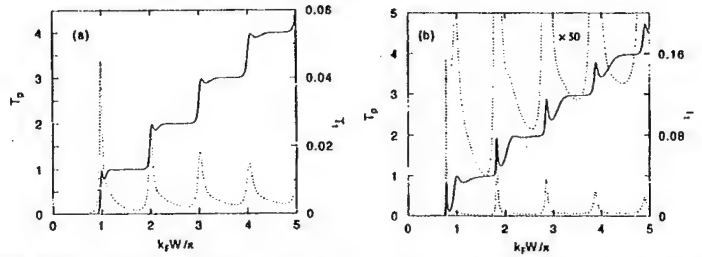


Figure 4. The tunneling probability through the point contact and the leakage probability into the side arm are shown as the Fermi energy is varied. In (a), k_Fd/π=0.1 and U/E_F=40, while in (b) k_Fd/π=1 and U/E_F=1.8. In both cases, k_FD=50 and k_FL=8.

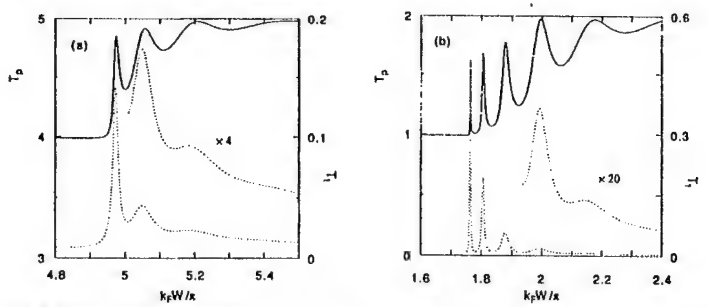


Figure 5. The transmission probability into the through and side arms for a long point contact with k_FD=50 and k_FL=8. In (a), k_Fd/π=0.1 and U/E_F=40, while in (b) k_Fd/π=1 and U/E_F=2.

Figure 4 shows the transmission characteristics as the Fermi energy is varied, which changes the number of occupied subbands. The tunneling probability, as plotted by the dottedt curves in the figure (and the next as well), is suppressed exponentially as the barrier potential, or its thickness, is increased. Peaks in the tunneling current line up with the steps in the in-line conductance, demonstrating that the tunneling peaks are measuring the 1D density of states in the quantum point contact region. In Fig. 5, the transmission probabilities are shown for long point contacts. As has been previously seen by others (Szafer and Stone, 1989; Kirczenow, 1989), a resonance structure due to multiple reflections from the two abrupt wide/narrow transitions in the waveguide is superimposed upon the conductance plateau. The resonances occur when $i\lambda_F=2L$, where i=1, 2, ..., and λ_F is the Fermi wavelength of an electron in the constriction. This resonance structure is replicated in the tunneling current as well.

Disorder-Induced Localization in Quantum Wires

In the preceding sections, quantum wires were discussed in terms of their modes of propagation, and it was noted that steps in the conductance can be observed. Unfortunately, most quantum wires are not sufficiently short to display good conductance quantization. Rather, inhomogeneities in the potential within the wire, whether due to random impurities (Nixon *et al.*, 1991) or to surface roughness (Thornton *et al.*, 1989; van Houten *et al.*, 1988; Akera and Ando, 1991), generally disrupt the coherence of the modal picture. This mixing of the modes provides loss of phase coherence within the guide, with the result that the sharp conductance steps are suppressed. In this section, we now turn to an example of mode matching using the scattering matrices as a method of treating this problem (Takagaki and Ferry, 1992c).

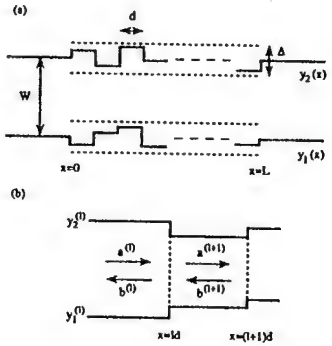


Figure 6. Schematic view of an electron waveguide with an inhomogeneous boundary. In each waveguide section with length d, the width is assumed to be uniform.

We model the problem by considering a waveguide with a rough boundary as shown in Fig. 6. An electron waveguide with a nonuniform width $w(x)=y_2(x)-y_1(x)$ is terminated by perfect leads with uniform width W. The wire structure is divided lengthwise into a few sections. An equivalent length d is assigned to each section, and the width in each section is taken to be uniform. In real devices, the boundaries $y_2(x)$ and $y_1(x)$ are smooth functions of the coordinate, but here we will take the length of each section smaller than any coherence length; that is, our model can simulate the arbitrary boundary contour if d is chosen sufficiently small compared to the Fermi wavelength λ_F. The approach to be taken is to make the following approximation. We introduce a local deviation in the width by defining $y_1=w_1$ and $y_2=W+w_2$, where w_1 and w_2 are uniformly distributed in the range $(-\Delta/2,\Delta/2)$. We also require that the two width variables are uncorrelated and

$$\langle w_1(x)\rangle = \langle w_2(x)\rangle = \langle w_1(x)w_2(x')\rangle = 0 \, , \tag{159}$$

and

$$\left\langle w_1(x)w_1(x')\right\rangle = \left\langle w_2(x)w_2(x')\right\rangle = \begin{cases} \dfrac{\Delta^2}{12}\left(1-\dfrac{|x-x'|}{d}\right), & |x-x'|<d, \\ 0, & |x-x'|>d \end{cases} . \tag{160}$$

For $d \ll \lambda_F$, the fluctuation resembles "white noise." The approximation in (160) is the first two terms of an exponential correlation function, which is seen both in the Si/SiO_2 interface (Goodnick *et al.*, 1985; Yoshinobu *et al.*, 1993) and the GaAs/AlGaAs interface (Feenstra, 1992).

The numerical technique to be followed is based upon extending the mode-matching technique to the scattering matrix approach. The wavefunction in the lth waveguide section $[(l-1)d<x<ld,\ y_{1l}<y<y_{2l}]$ is given by

$$\Psi_l(x,y) = \sum_j \frac{1}{\sqrt{k_{jl}}}\left[a_{jl}e^{ik_{jl}[x-(l-1)d]} + b_{jl}e^{-ik_{jl}(x-ld)}\right]u_{jl}(y) , \tag{161}$$

with

$$u_{jl}(y) = \sqrt{\frac{2}{y_{2l}-y_{1l}}}\sin\left[\frac{j\pi(y-y_{1l})}{y_{2l}-y_{1l}}\right] , \tag{162}$$

$$k_{jl} = \sqrt{k_F^2 - \left(\frac{j\pi}{y_{2l}-y_{1l}}\right)^2} , \tag{163}$$

for hard-walll boundary conditions. The summation runs over some practical upper limit, which includes an adequate number of evanescent modes.

For each junction between two waveguide sections, a scattering matrix is determined which relates the amplitude of incoming waves and outgoing waves. We define the scattering matrices $\mathbf{S}_l$ associated with the junction at $x=ld$ as (Cahay et al., 1988)

$$\begin{bmatrix} \mathbf{a}_{l+1} \\ \mathbf{b}_l \end{bmatrix} = \mathbf{S}_l \begin{bmatrix} \mathbf{a}_l \\ \mathbf{b}_{l+1} \end{bmatrix} = \begin{bmatrix} \mathbf{tr}' & \mathbf{a}_l \\ \mathbf{rt}' & \mathbf{b}_{l+1} \end{bmatrix} . \tag{164}$$

Here, $\mathbf{a}$ and $\mathbf{b}$ are column matrices representing the amplitude of the modes, and the matrices $\mathbf{t}$, $\mathbf{t}'$, $\mathbf{r}$, and $\mathbf{r}'$ describe the mode-mixing properties across the junction. The full scattering matrices have been worked out for this case by Takagaki and Ferry (1992c). These matrices are then cascaded together as

$$\mathbf{S} = \mathbf{S}_0 \otimes \mathbf{S}_1 \otimes \mathbf{S}_2 ..... \tag{165}$$

The overall conductance is then found from the multi-terminal Landauer-Büttiker formula (Landauer, 1957; Büttiker *et al.*, 1985)

$$G = \frac{2e^2}{h} Tr[\mathbf{tt}'] . \tag{166}$$

The reason for using the scattering matrix, rather than something like a transfer matrix approach, lies in the inherent stability of the former and its convenience in evaluating the net conductance. Moreover, the transfer matrix approach has stability problems for lengths longer than the Fermi wavelength.

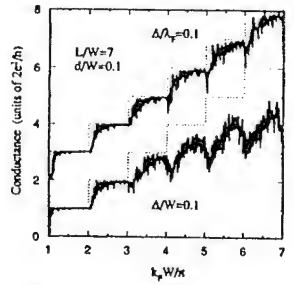


Figure 7. Energy dependence of the conductance of $L/W=7$ wires. The dotted lines represent the conductance of a perfect wire, and the upper curves have been offset for clarity.

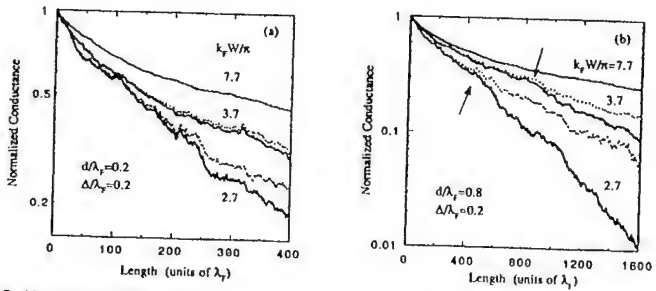


Figure 8. Normalized average conductance for a set of wires. In (a), the averaged conductance is shown as a function of length. For longer lengths (b), the dependence of the conductance on the length begins to show exponential localization behavior (arrows).

In Fig. 7, the energy dependence of the conductance is shown for a wire with $L/W=7$ and $d/W=0.1$. The conductance is shown for two different methods of normalizing the roughness, either as a fraction of the width ($\Delta/W=0.1$) or as a fraction of the Fermi wavelength ($\Delta/\lambda_F=0.1$). In the absence of the roughness, the conductance shows perfect steps. The roughness obscures the quantization. The electron states are essentially localized near the propagation thresholds, which leads to rounding of the steps. The fluctuations in the conductance arise from mode mixing and resonances within the individual laters. In Fig. 8, we show the length dependence of the conductance averaged over a group of wires ($g_{norm}=g/N$). The solid and dotted lines represent averages in terms of $\langle \ln(G)\rangle$ and $\ln\langle G\rangle$, respectively. There is no appreciable difference when the disorder is weak, but there is a significant difference in the strong disorder limit. One can see in panel (b) where there is a transition from power-law dependence on the length to exponential localization (arrows).

REFERENCES

Akera, H., and Ando, T., 1991, *Phys. Rev. B* 43:11676.

Bagwell, P. F., 1990, *Phys. Rev. B* 41:10354.

Balescu, R., 1975, "Equilibrium and Nonequilibrium Statistical Mechanics," Wiley, New York.

Barker, J. R., and Ferry, D. K., 1980a, *Solid-State Electron.* 23:519.

Barker, J. R., and Ferry, D. K., 1980b, *Solid-State Electron.* 23:531.

Bogoliubov, N. N., and Bogoliubov, N. N., Jr., 1982, "Introduction to Quantum Statistical Mechanics," World Scientific Press, Singapore.

Brunetti, R., Jacoboni, C., and Rossi, F., 1989, *Phys. Rev. B* 39:10781.

Büttiker, M., 1988a, *Phys. Rev. B* 38:9375.

Büttiker, M., 1988b, *Phys. Rev. B* 38:13297.

Büttiker, M., 1989, *Phys. Rev. B* 40:3409.

Büttiker, M., Imry, Y., Landauer, R., and Pinhas, S., 1985, *Phys. Rev. B* 31:6207.

Cahay, M., McLennan, M., and Datta, S., 1988, *Phys. Rev. B* 37:10125.

Feenstra, R. M., Yu, E. T., Woodall, J. M., Kirchner, P. D., Liu, C. L., and Pettit, G. D., 1992, *Appl. Phys. Lett.* 61:795.

Ferry, D. K., 1990, *in* "Granular Nanoelectronics," eds. D. K. Ferry, J. R. Barker and C. Jacoboni (Plenum, New York, 1990) p. 1.

Ferry, D. K., and Barker, J. R., 1980, *Solid-State Electron.* 23:545.

Frensley, W., 1990, *Rev. Mod. Phys.* 62:3.

Goldstein, H., 1959, "Classical Mechanics," Addison-Wesley, New York.

Goodnick, S. M., Ferry, D. K., Wilmsen, C. W., Lilienthal, Z., Fathy, D., and Krivanek, O. L., 1985, *Phys. Rev. B* 32:8171.

Han, J., Ferry, D. K., and Newman, P., 1990, *IEEE Electron. Dev. Lett.* 11:209.

Hillary, M., O'Connell, R. F., Scully, M. O., and Wigner, E. P., 1984, *Phys. Repts.* 106:121.

Iafrate, G. J., 1988, *in* "The Physics of Submicron Devices," Ed. by H. L. Grubin, D. K. Ferry, and C. Jacoboni, Plenum, New York, NATO ASI 180.

Ishibashi, A., Funato, K., and Mori, Y., 1988, *Jpn. J. Appl. Phys.* 27:L2382.

Kadanoff, L. P., and Baym, G., 1962, "Quantum Statistical Mechanics," Addison-Wesley, New York.

Kirczenow, G., 1989, *Phys. Rev. B* 39:10452.

Kumar, A., Laux, S. E., and Stern, F., 1988, *Appl. Phys. Lett.* 54:1270.

Landauer, R., 1957, *IBM J. Res. Develop.* 1:223.

Laux, S. E., Frank, D. J., and Stern, F., 1988, *Surf. Sci.* 196:101.

Mahan, G. D., 1990, "Many-Particle Physics," Plenum Press, New York.

Meijer, P. H. E., 1966, "Quantum Statistical Mechanics," Gordon and Breach, New York.

Morse, P. M., and Feshbach, H., 1953, "Methods of Theoretical Physics," McGraw-Hill, New York.

Nedjalkov, M., Dimov, I., Rossi, F., and Jacoboni, C., 1994, to be published.

Nixon, J. A., Davies, J. H., and Baranger, H. U., 1991, *Phys. Rev. B* 43:12638.

Ralls, K. S., Skocpol, W. J., Jackel, L. D., Howard, R. E., Fetter, L. A., Epworth, R. W., and Tennant, D. M., 1984, *Phys. Rev. Lett.* 52:228.

Rickayzen, G., 1980, "Green's Functions and Condensed Matter," Academic Press, New York.

Rossi, F., Poli, P., and Jacoboni, C., 1992, *Semicond. Sci. Technol.* 7:1017.

Rossi, F., Jacoboni, C., and Nedjalkov, M., 1994, *Semicond. Sci. Technol.* 9:580.

Ryan, J., Han, J., Kriman, A., Ferry, D. K., and Newman, P., 1989, *Solid-State Electron.* 32:1609.

Sols, F., Macucci, M., Ravaioli, U., and Hess, K., 1989a, *Appl. Phys. Lett.* 54:350.

Sols, F., Macucci, M., Ravaioli, U., and Hess, K., 1989b, *J. Appl. Phys.* 66:3892.

Szafer, A., and Stone, A. D., 1989, *Phys. Rev. Lett.* 62:300.

Ter Haar, D., 1961, *Rep. Prog. Phys.* 24:304.

Thornton, T. J., Roukes, M. L., Scherer, A., and van der Gaag, B. P., 1989, *Phys. Rev. Lett.* 63:2128.

Thouless, D. J., 1977, *Phys. Rev. Lett.* 39:1167.

Takagaki, Y., and Ferry, D. K., 1992a, *Phys. Rev. B* 45:6715.

Takagaki, Y., and Ferry, D. K., 1992b, *Phys. Rev. B* 45:12152.

Takagaki, Y., and Ferry, D. K., 1992c, *J. Phys. Cond. Matter* 4:10421.

van Houten, H., Beenakker, C. W. J., van Wees, B. J., and Mooij, J. E., 1989, *Surf. Sci.* 196:144.

Yoshinobu, t., Iwamoto, A., and Iwasaki, H., 1993, *in* "Proc. Solid State Device and Materials Conference, Chiba, 1993" p. 612.

QUANTUM CONFINED SYSTEMS: WELLS, WIRES, AND DOTS

U. Rössler

Institut für Theoretische Physik
Universität Regensburg
93040 Regensburg, Germany

REFLECTIONS ON TRANSPORT AND SYSTEM SIZES

The title of the ASI "Quantum transport in ultrasmall devices" provokes at the beginning reflections on transport and how it is influenced by the size of the systems, in which it is taking place. The transport of electric charge in all devices of our daily life (usually carried by electrons) does not raise such a question: all characteristic lengths which rule electrical conductivity as a material dependent quantity are much shorter than the system size of the devices. This is true even for the tiny transistors in Megabit-chips, where in Si technology channel lengths of 100 nm are nowadays standard in mass production. However, miniaturization and ultra-large-scale integration are still in progress and will sooner or later reach system sizes, where electron transport even in common devices becomes size dependent. In our laboratories, for especially prepared systems and under suitable conditions, this era has already begun and "quantum transport in ultrasmall devices" is a field of highly active research (Bate, 1988; Landauer, 1989; Ferry and Grubin, 1994).

Relevant Length Scales in Transport (Beenakker and van Houten, 1991)

Lengths, which are relevant for charge transport as material specific quantities and thus determine the electrical conductivity, can be divided into those which depend on scattering processes the charge carrier undergoes when travelling through the solid under the influence of an applied voltage, and those which result from the quantum mechanical nature of the carriers.

Scattering Lengths. In a perfect crystal at zero absolute temperature an electron subject to a periodic potential can be described as a Bloch wave, representing a modulated plane wave with reduced wave vector **k**; i.e., a coherent state with well-defined phase throughout the crystal. In real crystals the electron will be scattered from built-in imperfections of the crystal (point defects, dislocations) and thereby change its phase, yet in a deterministic way: the electron does not change its energy, the scattering process is elastic and reversible. The average distance covered between two such scattering events by a representative carrier of the system is the "elastic scattering length" or "elastic mean free path" l. (The scattering with impurities may eventually become inelastic, if it is connected with, e.g., an excitation of bound electrons.) Obviously, l does depend on the distribution of lattice imperfections in the sample and can be increased by improving the material quality. As the actual impurity configuration in a sample is not known, its impact on transport is considered by averaging over a distribution of impurity configurations (Abrikosov et al., 1965).

At finite temperature, scattering with phonons, which can be understood as thermal lattice imperfections, becomes possible and will at sufficiently high temperature dominate over the impurity scattering. This process will always be connected with a change of the energy (it is inelastic) and the change of the phase will be random: because of coupling to a heat bath the process is irreversible and destroys the coherence of the particle state. The average distance covered between two such processes is the "inelastic" L_{in}. Clearly L_{in} depends on the temperature T, it can be much larger than l for sufficiently low temperature.

Some care is required concerning the inelastic processes because of a finite lifetime of the carrier, which leaves its energy uncertain by an amount ΔE. Therefore, inelastic scattering processes are essentially those which change the particle energy by more than ΔE.

Quantum Mechanical Lengths: In quantum (wave) mechanics a wavelength is ascribed to a particle with kinetic energy $\mathcal{E}$, the de Broglie wavelength $\lambda = h/(2m\mathcal{E})^{1/2}$. λ defines the length scale at which a diffracting obstacle will disclose the wave-like, i.e.~quantum mechanical, nature of the particle.

According to Pauli's principle the electrons in a metal or degenerate semiconductor at $T = 0$ K fill all states up to the Fermi energy $\mathcal{E}_F$ or in k-space the Fermi sphere with radius k_F. The quantum mechanical wave function of a particle with energy $\mathcal{E}_F$ has the typical wavelength $\lambda_F = 2\pi\ k_F^{-1}$ (Fermi-wavelength). The particles at the Fermi edge are essential for transport properties, because they can receive small amounts of energy in an accelerating electric field or by scattering processes to reach an empty state above $\mathcal{E}_F$. At finite temperature, the Fermi edge is smeared out and particles in an energy interval k_BT (k_B being the Boltzmann constant) around E_F take part in transport processes. Obviously λ_F (or k_F) depends on the electron density, which is orders of magnitudes larger for metals than for degenerate semiconductors.

As we will see in what follows, transport properties are frequently, and with tremendous success, investigated in the presence of a magnetic field. The oscillating motion of a mobile charge carrier in a magnetic field is characterized by the magnetic length $l_B = (\hbar/eB)^{1/2}$, which is the natural unit of length for this quantum mechanical problem (the corresponding unit of energy is the cyclotron energy $\hbar\omega_c = \hbar eB/m$. The importance of the magnetic length and its relevance for transport phenomena lies in the fact, that it can be tuned over a wide range from the outside by changing the magnetic field.

Transport Regimes

The length scales introduced above vary independently due to their specific dependence on temperature, particle density, impurity concentration, kinetic energy, or applied magnetic field and may eventually compete with each other or with the size of the system L in one or the other direction. This allows us to distinguish different transport regimes.

Classical Diffusive Transport. This marks the regime in which all electronic devices of our daily life operate. The system size L is irrelevant and conductivity can be

described by a drifting electron motion with successive acceleration in the direction of the applied electric field E and elastic scattering after a time τ out of this direction. The drift velocity is given by

$$\mathbf{v}_d = -\mu\mathbf{E} \ , \tag{1}$$

with the mobility $\mu = e\tau/m$ and the current density parallel to **E** is the charge density $-en$ moving with this drift velocity. Thus we arrive at the Drude expression for the static conductivity (Ashcroft and Mermin, 1976)

$$\sigma = \frac{ne^2\tau}{m} . \tag{2}$$

Fermi level properties (and thus the Pauli principle) enter the problem when we consider the balance between the drift current density and the diffusion current density due to gradients of the carrier density in thermodynamic equilibrium (i.e.~for constant electrochemical potential $\phi = \mathcal{E}_F - eV$)

$$-\frac{1}{e}\sigma\mathbf{E} - D\nabla^2 n = 0 , \quad \text{for } \nabla\phi = 0 . \tag{3}$$

with the diffusion constant D. Using $\mathbf{E} = -\nabla\phi$ and

$$\frac{dn}{d\mathcal{E}}\Big|_{\mathcal{E}_F} = \rho(\mathcal{E}_F) \ ,$$

for T = 0 K, we may write

$$\nabla\phi = e\mathbf{E} - \frac{1}{\rho(\mathcal{E}_F)}\nabla n \ , \ . \tag{4}$$

where $\rho(\mathcal{E}_F)$ is the density of states at $\mathcal{E}_F$. Comparison of (1.3) with (1.4) yields the Einstein relation (Ashcroft and Mermin, 1976)

$$\sigma = e^2\rho(\mathcal{E}_F)D \quad . \tag{5}$$

The relations between carrier density and Fermi energy or Fermi wavelength depend on the dimensionality of the system. The diffusive regime introduces another length, the coherence length $L_\phi = (D\tau_\phi)^{1/2}$, which in general is different from L_{in}.

Coherent Transport ($L \le L_\phi$). For system sizes L smaller than the coherence length L_ϕ we meet a situation where the quantum mechanical wave function of the charge carrier has a well defined phase all over the system (say from one lead to the other one). In this situation transport will show quantum interference phenomena (Aharonov-Bohm oscillations, universal conductance fluctuations; see lectures by L. Eaves in this volume) which can be taken as fingerprints of the individual sample and its impurity distribution.

After a series of elastic scattering processes with impurites a carrier may eventually return to its starting point (coherent backscattering; see Bergman, 1983), i.e.~a destructive interference of probability amplitudes along different paths becomes possible. This is the mechanism behind the phenomenon of weak localization (Anderson *et al.*, 1979). It will take place irrespective of the system size if $l >> \lambda_F$ and is well known in disordered systems.

Ballistic transport ($L \le l$): When the system size L gets smaller than the elastic mean free path (and at sufficiently low temperatures that $L_\phi > l$), a carrier can cross the device without any scattering in analogy to electromagnetic waves in a waveguide. In this case, transport is a transmission problem which has been studied in point contacts and constrictions (see seminars by Jacoboni and Barker) and is well understood within the concepts of Landauer (1957, 1988) and Büttiker (1986).

Size Quantization ($L \approx \lambda$). As was already mentioned the de Broglie wavelength defines the length scale on which an obstacle makes the wave nature of a particle obvious. This obstacle can be boundaries of the system which confine the particle motion or else allow only for standing wave solutions, i.e.~bound states with discrete energies in a potential. This size quantization, which is connected with a change in the density of states, can be realized by confinement in 1, 2 or 3 space directions and will lead to quantum wells, wires, and dots, respectively. This aspect of dimensionality is the subject of the next section.

Table 1. Characteristic electron parameters for a two-dimensional electron gas (2DEG).

Density, n	1-10	10^{11} cm^{-2}
Fermi wavelength, λ_F	25-80	nm
Parameters which depend upon m^*=0.066m_0:		
Fermi energy, $\mathcal{E}_F$	3.6-36	meV
Fermi velocity, v_F	1.4-4.3	10^7 cm/s
Parameters, which depend on mobility $\mu = 10^6$ cm^2/V-s:		
Elastic scattering time, τ	40	ps
Mean free path, $l = v_F\tau$	5-16	μm

In Table. 1, we have collected numerical values for the quantities introduced in this section for two-dimensional electrons in GaAs heterostructures as a representative system. In contrast to the Fermi wave length, which is material independent and uniquely determined solely by the density, the Fermi energy and Fermi velocity $v_F = \hbar k_F/m$ also depend on the effective mass of the carriers. The mobility (at low temperatures) is determined by elastic impurity scattering and gives an elastic scattering time τ that is independent of n. The coherence length L_ϕ (in comparison to l) depends on the transport regime as discussed.

Dimensionality

The principle quantum mechanical aspects of confined carriers can be discussed already using the picture of a particle in the box. As we shall see frequently during this

seminar, layered semiconductor systems eventually combined with lateral structuring allow us to realize what seems to be highly simplified models in basic quantum mechanics. This led one of the pioneers in this field, L. Esaki (1991), to talk about "do-it-yourself quantum mechanics".

Given the kinetic energy

$$\mathcal{E}(k) = \frac{\hbar^2 k^2}{2m} \quad , \quad k^2 = k_x^2 + k_y^2 + k_z^2 \quad , \tag{6}$$

of electrons in the conduction band of a simple metal or semiconductor (in the latter case $m = m^*$, the effective mass, which is unusually much smaller than the free electron mass m_0) we count the states by assuming confinement to a box with side lengths L_x, L_y, L_z. As a consequence the values of k_x, k_y, k_z are virtually quantized with discrete differences $\Delta k_i = \{2\pi/L_i\}$ $i = x,y,z$. This allows us to calculate the number of states up to an energy $\mathcal{E}$ by dividing the volume of a sphere with radius $k(\mathcal{E}) = (2m\mathcal{E}/\hbar)^{1/2}$ by the volume alloted to each **k**. If spin degeneracy is included by a factor of 2 we have

$$N(\mathcal{E}) = 2\,\frac{L_x L_y L_z}{8\pi^3}\,\frac{4\pi}{3}k^3 \tag{7}$$

as the number of states in the box up to the energy $\mathcal{E}$. After division by the volume of the box, $L_x L_y L_z$, the number

$$n_{3D}(\mathcal{E}) = \frac{N(\mathcal{E})}{L_x L_y L_z} = \frac{1}{3\pi^2}k^3(\mathcal{E}) \quad , \tag{8}$$

refers to the unit volume of the material. The density of states is obtained as the derivative $\rho(\mathcal{E}) = dn(\mathcal{E})/d\mathcal{E}$, i.e.~for the 3D-case considered here (see Fig. 1a),

$$\rho_{3D}(\mathcal{E}) = \frac{1}{2\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2}\mathcal{E}^{1/2} \quad . \tag{9}$$

It is important to recall that the box dimensions are irrelevant in this case and the box is used only as a trick for counting states in a continuum. The expressions in (1.8) and (1.9) are valid also in the limit of an infinite box.

If we consider confined systems, the box becomes reality by the influence it plays on the spectrum. For a quantum well of width L_z grown in the z-direction, we have to consider instead of (1.6), the following

$$\mathcal{E}_\nu(k) = \mathcal{E}_\nu + \frac{\hbar^2 k^2}{2m} \quad , \quad k^2 = k_x^2 + k_y^2 \quad , \tag{10}$$

with discrete energies

$$\mathcal{E}_\nu = \frac{\hbar^2}{2m}\left(\frac{\nu\pi}{L_z}\right)^2 \quad ,$$

due to size quantization. Each $\mathcal{E}_\nu$ defines the bottom of a 2D subband with respect to dispersion in k_x and k_y direction. Counting the states in a subband from $\mathcal{E}_\nu$ up to an energy $\mathcal{E}$, we have now to divide the area of a circle with radius $k(\mathcal{E})$ by the area ascribed to each 2D **k** vector. Considering spin degeneracy and dividing by $L_x L_y$, we find for the individual subband

$$n_{2D}(\mathcal{E}) = \frac{1}{2\pi}k^2(\mathcal{E}) = \frac{m}{\pi\hbar^2}\,\mathcal{E} \quad , \tag{11}$$

and the density of states of a series of subbands is

$$\rho_{2D}(\mathcal{E}) = \frac{m}{\pi\hbar^2}\sum_\nu \Theta(\mathcal{E}-\mathcal{E}_\nu) \quad . \tag{12}$$

Each time the energy reaches a new subband the density of states makes a jump by $m/\pi\hbar^2$ (see, for example, Fig. 1b).

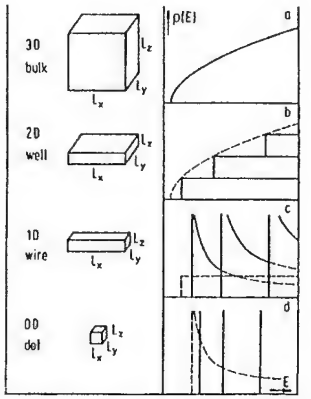


Figure 1. Geometry, box lengths and density of states for a) 3D, b) 2D, c) 1D, and d) 0D systems.

Confinement in two space directions, say y and z, gives a spectrum (Fasol et al., 1989; Beaumont et al., 1990; Rosenbacher et al., 1992)

$$\mathcal{E}_{\nu,\mu}(k_x) = \mathcal{E}_{\nu,\mu} + \frac{\hbar^2 k^2}{2m} \quad , \quad \mathcal{E}_{\nu,\mu} = \frac{\hbar^2}{2m}\left[\left(\frac{\nu\pi}{L_z}\right)^2 + \left(\frac{\nu\pi}{L_y}\right)^2\right] \quad , \tag{13}$$

i.e.~subbands with dispersion merely in one direction (quantum wires). The number of states up to an energy $\mathcal{E}$ above a given $\mathcal{E}_{\nu\mu}$ and per unit length is

$$n_{1D}(\mathcal{E}) = \frac{2}{\pi}k(\mathcal{E}) = \frac{2}{\pi}\left(\frac{2m\mathcal{E}}{\hbar^2}\right)^{1/2} \quad , \tag{14}$$

leading to density of states (see Fig. 1c)

$$\rho_{1D}(\mathcal{E}) = \frac{1}{\pi}\left(\frac{2m}{\hbar^2}\right)^{1/2} \sum_{\nu\mu} (\mathcal{E} - \mathcal{E}_{\nu,\mu})^{-1/2}\Theta(\mathcal{E} - \mathcal{E}_{\nu,\mu}) \quad . \tag{15}$$

Finally, for confinement in all three space directions, we have quantum dots with a completely discrete spectrum

$$\mathcal{E}_{\nu,\mu,\lambda} = \left(\frac{\hbar^2}{2m}\right)\left[\left(\frac{\nu\pi}{L_z}\right)^2 + \left(\frac{\mu\pi}{L_y}\right)^2 + \left(\frac{\lambda\pi}{L_x}\right)^2\right] \quad , \tag{16}$$

and the density of states becomes a series of δ-peaks

$$\rho_{0D}(\mathcal{E}) = 2\sum_{\nu\mu\lambda} \delta(\mathcal{E} - \mathcal{E}_{\nu,\mu,\lambda}) \quad , \tag{17}$$

as shown in Fig. 1d.

Each of the confinement lengths L_x, L_y, L_z can separately get in competition with the lengths introduced above, thus disclosing the dimension of the system. According to (8), (11) and (14) (used for $\mathcal{E} = \mathcal{E}_F$) the Fermi wavelength (at $T = 0$ K)

$$\lambda_F = 2\pi k_F^{-1} \sim n^{-1/d} \tag{18}$$

depends on the particle density n differently for d =1, 2, 3 dimensional systems.

Let us consider now a homogeneous magnetic field **B** applied in the z-direction. It quantizes the otherwise continuous spectrum as described by replacing in (6) or (10) $(\hbar^2/2m)(k_x^2+k_y^2)$ by $\hbar\omega_c(n+1/2)$, thus obtaining the spectrum

$$\mathcal{E}_N(k_z) = \hbar\omega_c(N+1/2) + \frac{\hbar^2 k^2}{2m} \quad . \tag{19}$$

for a 3D system and

$$\mathcal{E}_{N,\nu} = \mathcal{E}_\nu + \hbar\omega_c(N+1/2) \tag{20}$$

for a 2D system. Here $N = 0, 1, 2,...$ is the Landau quantum number and $\hbar\omega_c = \hbar eB/m$ is the cyclotron energy. The density of states corresponding to (19) and (20) has the same structure as those obtained for the 1D (Fig. 1c) or 0D system (Fig. 1d) but now with equally spaced peaks separated by $\hbar\omega_c$. The Landau levels are highly degenerate, with a degeneracy factor obtained by dividing the system size L_xL_y in the plane perpendicular to **B** by $2\pi l_B^2$, the area enclosed by a cyclotron orbit. Per unit area (after division by L_xL_y, we find

$$n(B) = \frac{2eB}{h} \quad , \tag{21}$$

where the factor 2 counts the two spin directions. This degeneracy results from the translational invariance of the system in the (x,y)-plane perpendicular to **B**. If it is removed by impurities or by lateral structure, i.e.~formation of quantum wires or dots, the degeneracy will be lifted.

ABOUT BARRIERS AND WELLS

With the advent of the planar Si technology for large scale integration in the 1960's, interest in low dimensional semiconductor structures has experienced a dramatic increase. Growth techniques, which allow a controlled layer by layer composition of such structures have been developed and refined. Nowadays the dominating ones are molecular beam epitaxy (MBE) (Tsao, 1993) and metal-organic chemical vapour deposition (MOCVD) (Duchemin *et al.*, 1985). Besides the fundamental idea to realize artificial periodic structures with lattice constants smaller than the mean free path l (Esaki and Tsu, 1970), there was also the device oriented expectation to get a hand on the material properties (band structure engineering) (Narayanamurti, 1984; Capasso, 1987). The inhomogeneity of the layered semiconductor structure in growth direction (with layer thickness L_z large compared with the atomic lattice constant of a few tenths of nanometers leads to piecewise constant conduction and valence band edges as indicated in Fig. 2. They will be experienced by particles (elecrons in the conduction band, holes in the valence band) as barriers, superlattices or wells with different profiles (see Fig. 2). Barriers and likewise superlattices are interesting objects for transport experiments in the ballistic regime, when the mean free path gets larger than the characteristic length of the barrier structure (Tsu and Esaki, 1973; Chang *et al.*, 1974). The characteristic feature of wells, the quantum size effect, becomes important when the well width is of the order of the de Broglie wavelength λ (Dingle *et al.*, 1974).

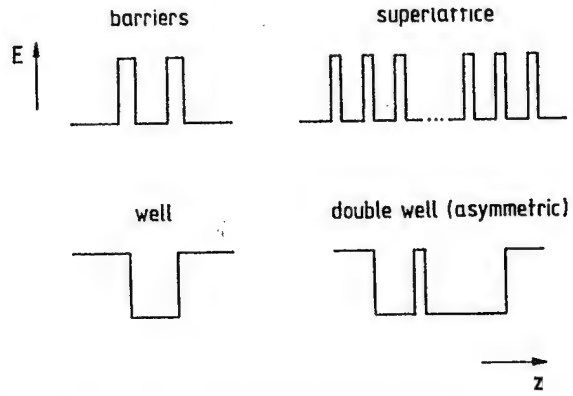


Figure 2. Potential profiles for different types of layered semiconductor structures.

Transport Through Barriers

Elementary quantum mechanics provides all the concepts to describe transport of a particle through a single barrier or a series of periodic barriers by solving the 1-dimensional Schrödinger equation with piecewise constant potential. Tunneling through n identical barriers can be described by transmission (T) and reflection (R) amplitudes connected by a sequence of n transfer matrices for the individual barriers (Tsu and Esaki, 1973)

$$\begin{pmatrix} T \\ 0 \end{pmatrix} = M_n \ldots M_p \ldots M_1 \begin{pmatrix} 1 \\ R \end{pmatrix} = \mathrm{M}\begin{pmatrix} 1 \\ R \end{pmatrix} \quad , \tag{22}$$

where M is a unitary 2×2 matrix. From (22), we obtain, using $T = M_{11} + M_{12}$ and $D = M_{21} + M_{22}$, the transmission amplitude is

$$T = M_{11} - \frac{M_{12}M_{21}}{M_{22}} \quad . \tag{23}$$

The matrix elements of M depend on the geometry and height of the barrier and of the particle energy $\mathcal{E}$.

The actual transport problem has to include the voltage applied over the sequence of barriers (which defines the length L of our system) and leads to a net current density through the barriers

$$j = \frac{e}{4\pi^3\hbar}\int_0^\infty dk_z \int_0^\infty dk_t \; 2\pi k_t [f(\mathcal{E})-f(\mathcal{E}')]T^*T\frac{\partial \mathcal{E}}{\partial k_z} \quad . \tag{24}$$

where $\mathcal{E}$ ($\mathcal{E}'$) is the energy of the incident (transmitted) electron, which has been split into transverse (parallel) and longitudinal z contributions corresponding to projections of the wave vector parallel to the interface planes k_t or along the growth direction k_z. $f(\mathcal{E})$ and $f(\mathcal{E}')$ are the Fermi distributions for emitter and collector, respectively. In the low temperature limit, (24) can be evaluated and gives

$$j = \frac{em^*}{2\pi^2\hbar^3}\int_0^{E_F} d\mathcal{E}_L \, [\mathcal{E}_F-\mathcal{E}_L] \, TT^* \qquad , \; V > \mathcal{E}_F \; , \tag{25a}$$

$$j = \frac{em^*}{2\pi^2\hbar^3}\Big\{ \int_0^{E_F-V} d\mathcal{E}_L \, TT^* + \int_{E_F-V}^{E_F} d\mathcal{E}_L \, [\mathcal{E}_F-\mathcal{E}_L] \, TT^* \qquad , \; V < \mathcal{E}_F \; . \tag{25b}$$

In its dependence on the voltage V across the sample $ln(j)$ shows pronounced peaks which, however, in the first successful experiments on a 80 nm $Al_{0.7}Ga_{0.3}As$/50 nm GaAs/80 nm $Al_{0.7}Ga_{0.3}As$ double barrier structure could be detected only in the derivative dj/dV vs. V (Chang *et al.*, 1974). It turns out that the transmission becomes particularly efficient under resonance conditions with the quantized states in the inner well (resonant tunneling). The proposal (Esaki and Tsu, 1970, 1973) and realization (Chang *et al.*, 1974) of tunneling through barriers in layered semiconductor structures has been a particularly clever realization of simple quantum mechanics.

Since then, major improvements in sample quality have led to observations of negative differential resistance up to room temperature and to a tremendous increase in the peak to valley ratios in the currents through double-barrier structures as a basis for device applications in resonant tunneling diodes (Solner et al., 1991) (Fig. 3) and hot electron transistors (Heiblum et al., 1987).

In p-type material, resonant tunneling through double barrier structures is more complex due to the valence band structure and the heavy-light-hole mixing at finite in-plane wave vector k_t. Besides the demonstration of the existing two types of holes (Mendez *et al.*, 1977), resonant tunneling with an in-plane magnetic field has been measured to probe the complex hole subband dispersion (Hayden et al., 1991). The $I(V)$ characteristics obtained for different magnetic fields B in a setup, shown schematically in Fig. 4a, exhibit peaks at voltages which vary with B (Fig. 4b). By treating the in-plane magnetic field as a perturbation, these authors ascribe to each B value an in-plane wave vector k_t and transform the applied voltage to an energy scale to obtain dispersion curves for hole subbands, which resemble calculated ones from simplified models. We come back to this point in the next section.

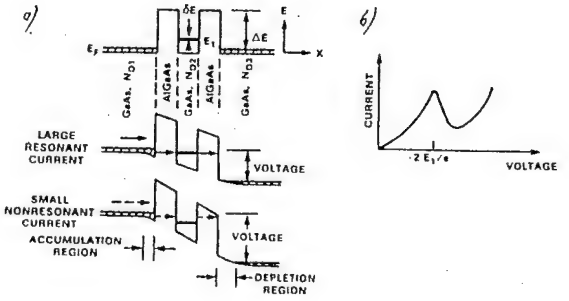


Figure 3. Resonant tunneling diode: a) energy diagram without and with applied voltage, b) current-voltage characteristic (from Sollner *et al.*, 1991).

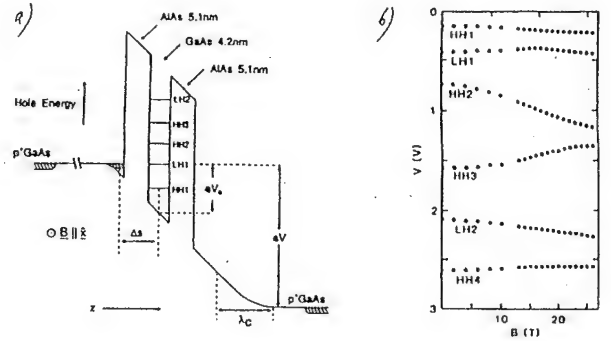


Figure 4. a) Double barrier structure used for resonant magneto tunneling experiments, b) peaks read from current-voltage characteristics for different magnetic field applied along z (from Hayden *et al.*, 1991).

Resonant tunneling through barriers has been used also to demonstrate quantization in dot structures (Reed *et al.*, 1989). Such structures were fabricated by etching techniques and are sketched schematically in Fig. 5 together with their energy diagram. The peak structure observed at low temperatures in the $I(V)$ characteristics was taken as evidence of the level quantization, which according to the size of the dots is about 1 meV. For more details and results of higher actuality I refer to Mark Reeds lectures in this volume.

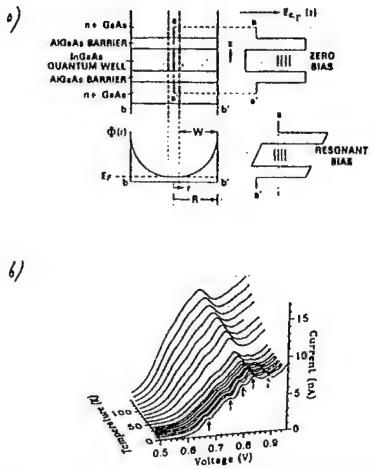


Figure 5. Schematic view of tunneling through a dot structure (a) and current voltage characteristics for different temperatures (b) (from Reed *et al.*, 1989).

The study of electron wave dynamics in periodic potentials in the early days of quantum mechanics (Bloch, 1928) has led to the prediction of Bloch oscillations: a Bloch electron in a potential with period d and a uniform electric field F should oscillate with a frequency $\nu_B = (eFd/h)$, due to Bragg reflections at the Brillouin zone edge. Observation of Bloch oscillations is possible only in the ballistic regime, i.e.~the elastic scattering time τ should not be smaller than the period $1/\nu_B$ of the Bloch oscillation. It turned out, that due to the small lattice constants unrealistic high fields had to be applied to meet such a condition in conventional solids. However, as pointed out already by Tsu and Esaki (1973), semiconductor superlattices should do it. Even though the evidence of Bloch oscillations has long been searched for and debated with some controversy (Mendez, 1993; Feldman, 1992), because of the competing effect of the Wannier-Stark ladders. Only recently, Bloch oscillations have been directly detected by the coherent electromagnetic radiation emitted from the periodically accelerated electron (Waschke *et al.*, 1993).

Quantum Wells

The quantum size effect in quantum wells has been observed first in optical experiments on multiple quantum wells (Dingle *et al.*, 1974; Dingle, 1975) at about the same time as the resonant tunneling through a sequence of barriers. Optical transmission (or absorption) through quantum wells of width decreasing from L_z = 400 nm down to less than 10 nm exhibit a change from the bulk like spectrum to one which shows the characteristic onsets of excitation between pairs of different hole and electron subbands (Fig. 6). The blue shift of the onset demonstrates the increase of the confinement energy with decreasing L_z --- an effect which is used in optoelectronic applications to fit the effective energy gap to a wavelength at which optical fibers connected with the quantum well device have ideal transmission properties (band structure engineering). The increase of the subband separations with decreasing L_z is reflected in an increasing separation of the structures dominating in the spectra. In a simplified single particle picture the absorption is given by

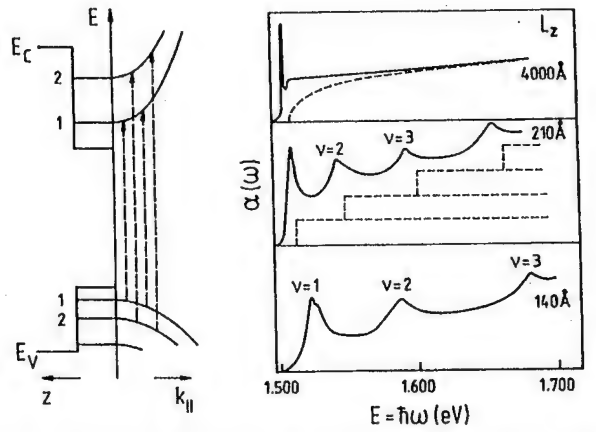


Figure 6. Intersubband transitions (left) and optical absorption spectra for quantum wells of different width L_z (right). The dashed lines indicate single-particle combined density of states (after Dingle *et al.*, 1974; Dingle, 1975).

$$\alpha(\omega) \sim \sum_{\substack{\nu_e \nu_h \\ \mathbf{k}_\parallel}} |M_{c\nu_e\ \nu\nu_h}|^2\, \delta(\mathcal{E}_{c\nu_e}(\mathbf{k}_\parallel) - \mathcal{E}_{\nu\nu_h}(\mathbf{k}_\parallel) - \hbar\omega) \tag{26}$$

with the subband dispersion

$$\mathcal{E}_{n\nu}(\mathbf{k}_\parallel) = \mathcal{E}_\nu + \mathcal{E}_n + \frac{\hbar^2 k_\parallel^2}{2m_n} \quad . \tag{27}$$

Here $\mathcal{E}_n$ ($n = v,c$)$ denote the bulk band edge energies of valence and conduction band. The dipole matrix element can be written

$$M_{c\nu_e\ \nu\nu_h} = \int dz\, \xi_c(\nu_e, z)\, \xi_c(\nu_h, z) \langle c|e \bullet p|v\rangle \quad , \tag{28}$$

with the subband functions $\xi(\nu, z)$ and the dipole matrix element between band edge Bloch functions. As $M_{cv}(\nu_c,\nu_h)$ does not depend on the in-plane wave vector $\mathbf{k}_\parallel$ the sum over this quantity with the δ-function gives the step-like function known for the density of states above. The integral over the subband functions is essentially different from zero only for $\nu_e = \nu_h$ (this is exact in an infinite barrier model). Thus $\alpha(\omega)$ is described as the combined density of states of electron and hole subbands with same index ν.

Deviations of the experimental results from this simple picture indicate the complexity of the real system due to band structure effects and electron-hole correlation (excitonic effects). These effects shall be considered in this section.

Band Structure Effects. The envelope function approximation (EFA) is the standard concept to describe subband states. It is based on the fact that these states cover an energy range of a few hundreds of meV above the bulk band edges and are determined essentially of this part of the bulk band structure. This concept has been developed and refined over the last years (Bastard et al., 1991). For standard semiconductors like GaAs, the conduction band minimum (symmetry Γ_6) derives from *s*-antibonding states, its dispersion to lowest order in **k** reads

$$\mathcal{E}_c(\mathbf{k}) = \mathcal{E}_c(\Gamma_6) + \frac{\hbar^2 k^2}{2m_c} \ , \ \mathbf{k} = (k_x, k_y, k_z) \ , \tag{29}$$

and is twofold spin-degenerate. Replacing $k_z \rightarrow -i(\partial/\partial z)$, we obtain the kinetic energy operator of the effective mass approximation. For the subband problem we have to consider the inhomogeneity of the system by taking into account a *z*-dependent subband edge $\mathcal{E}_c(\Gamma_6)$, which can formally be written as the quantum well potential $V(z)$: the subband Hamiltonian reads

$$H_c = \frac{\hbar^2}{2m_c}\frac{\partial^2}{\partial z^2} + V(z) \ . \tag{30}$$

The solutions are:

$$\psi_v(x,y,z) = \xi_v(z)\exp[i(k_x x + k_y y)] \ , \ \mathcal{E}_v(\mathbf{k}_t) = \mathcal{E}_v + \frac{\hbar^2 k_t^2}{2m_c} \ . \tag{31}$$

Actually, also the effective mass m^* has to be considered as *z*-dependent because it is different in well and barrier.

For holes, the valence band structure deriving from *p*-bonding states (symmetry Γ_8) is obtained by diagonalizing the Luttinger Hamiltonian (Bastard, 1991; Andreani et al., 1987; Goldoni and Fasolino, 1992; Winkler and Rössler, 1993, 1994)

$$H_{4\times4} = \begin{bmatrix} P+Q & S & R & 0 \\ S^+ & P-Q & 0 & R \\ R^+ & 0 & P-Q & -S \\ 0 & R^+ & -S^+ & P+Q \end{bmatrix} , \tag{32}$$

where

$$P \pm Q = \mathcal{E}_v + \frac{\hbar^2}{2m_0}\left\{[\Gamma_1 \mp \Gamma_2]k_z^2 - [\Gamma_1 \pm \Gamma_2]\,k_t^2\right\}$$
$$R = -\frac{\sqrt{3}\hbar^2}{2m_0}\left\{\overline{\Gamma}k_-^2 + \mu k_+^2\right\} \ , \ S = -\frac{\sqrt{3}\hbar^2}{2m_0}k_z k_- \ , \tag{33}$$

and $k_\pm = k_x \pm ik_y$, $\overline{\Gamma} = (\Gamma_2 + \Gamma_3)/2$, $\mu = \Gamma_2 - \Gamma_3$. The bulk valence band is fourfold degenerate at $\mathbf{k} = 0$ and splits into two spin-degenerate parabolic bands whose curvature depends on the direction of **k** (warping).

The hole subband dispersion is obtained again by replacing $k_z \rightarrow -i(\partial/\partial z)$ and introducing *z* dependent valence band edge energies $\mathcal{E}_v(\Gamma_8,z)$ and quantum well potential $V_h(z)$ for holes. At $\mathbf{k}_t = 0$, the matrix Hamiltonian falls into pairwise identical diagonal parts

$$H_\pm = \frac{\hbar^2}{2m_0}[\Gamma_1 \mp \Gamma_2]\frac{\partial^2}{\partial z^2} + V_h(z) \ , \tag{34}$$

with two ladders of subband states for heavy holes (*hh*, upper sign) and light holes (*lh*, lower signs) with energies $\mathcal{E}_{hh}(\nu_h)$ and $\mathcal{E}_{lh}(\nu_l)$ respectively. The lowest heavy hole state has a smaller confinement energy than the light hole state due to the larger mass. Taking into account only the lowest *lh* and *hh* subband state, we may derive their in-plane dispersion as a solution of a 2×2 matrix

$$\begin{bmatrix} W_{hh} & R(\mathbf{k}_t) \\ R(\mathbf{k}_t) & W_{lh} \end{bmatrix} \tag{35a}$$

where

$$W_{hh} = \mathcal{E}_{hh1} - \frac{\hbar^2}{2m_0}[\Gamma_1+\Gamma_2]\,k_t^2 \ , \ W_{lh} = \mathcal{E}_{lh1} - \frac{\hbar^2}{2m_0}[\Gamma_1+\Gamma_2]\,k_t^2 \ . \tag{35b}$$

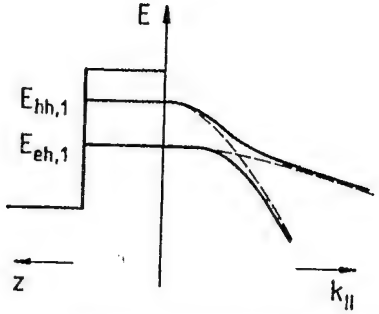


Figure 7. Quantum well and hole subband dispersion with (solid lines) and without (dashed lines) heavy hole- light hole coupling.

Without regard to the off-diagonal term we find parabolic dispersion with a larger (smaller) curvature for the heavy (light) hole states (see dashed lines in Fig. 7) which cross each other at finite k_t. Taking the off-diagonal terms into account, level repulsion takes place and the subband dispersion shows strong nonparabolicity (solid lines in Fig. 7). Thus two band structure effects are characteristic for holes in quantum wells: the lower symmetry of the layered system removes the fourfold degeneracy of the topmost valence band (Γ_8) allowed in the cubic symmetry of the bulk (*lh*-*hh* splitting), the finite in-plane wave vector mixes heavy and light hole states and causes a highly nonparabolic subband dispersion. Both effects depend on the band offsets, the Luttinger parameters and on L_z. In addition biaxial strain due to lattice mismatch takes influence on the *lh*-*hh* splitting.

Calculations of the hole subband dispersion in a rectangular quantum well can be carried out analytically using the 4×4 Luttinger Hamiltonian (Andreani *et al.*, 1987). More

complex potential profiles, like asymmetric double quantum wells (see Fig. 2), reveal the removal of spin-degeneracy (Goldoni and Fasolino, 1992). It turns out that even the coupling to the spin-split-off band (symmetry Γ_7 in the bulk band structure) and to the conduction band influences the hole subband structure. These aspects have be considered by using multiband concepts based on **k•p** Hamiltonians for up to 8 bands, including the Γ_6 conduction and the Γ_8 and Γ_6 valence bands, whose solutions are found by transformation into k-space, thus solving instead of 8 coupled differential equations the corresponding integral equations using quadrature methods (Winkler and Rössler, 1993, 1994). Results of these calculations for AlAs/GaAs quantum wells with L_z = 4.2 nm and 6.8 nm are shown in Fig. 8 for a 4×4 (Γ_8, dotted lines), a 6×6 (Γ_8 and Γ_7, dashed lines) and the full 8×8 model (Γ_6, Γ_8 and Γ_7, solid lines). These results demonstrate that for quantitative interpretations of data for hole subbands in GaAs quantum wells the split-off valence band can not be neglected. Using these results in connection with self-consistent calculations for the double-barrier structure of Hayden *et al.* (1991) raises some doubt in the interpretation of the resonant magneto-tunneling data (Winkler and Rössler, 1993, 1994).

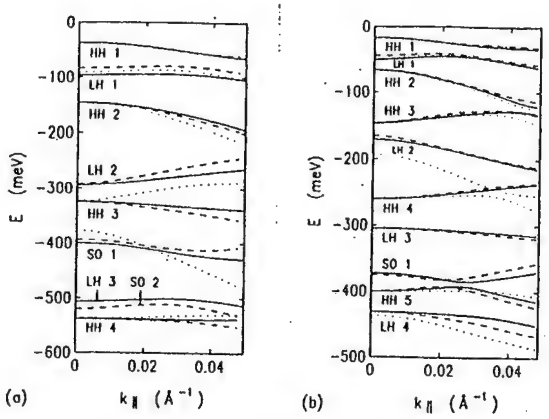


Figure 8. Hole subband dispersion for AlAs/GaAs quantum wells with L_z = 4.2 nm (a) and 6.8 nm (b) calculated from different multiband models: 4×4 (dotted lines), 6×6 (dashed lines), 8×8 (solid lines) (from Winkler and Rössler, 1993, 1994).

Excitonic effects. The absorption constant of bulk semiconductors is known to be strongly modified by excitonic effects (Reynolds and Collins, 1981; Rössler, 1979). In quantum well structures these effects are enhanced, because quantum confinement localizes both electron and hole to the quantum well, thus increasing the electron-hole Coulomb interaction. This is already seen in Dingle's (1975) experiments, which exhibit peaks instead of the steplike absorption structures as described in the single particle picture. More recent experiments resolve more details of these excitonic effects (Dawson et al., 1986; Kajikawa, 1993). The electron-hole correlation can be considered in EFA by writing the exciton envelope function as

$$\psi_\sigma(r_e,r_h) = \frac{1}{4\pi^2}\sum_{j_e j_h}^{v_e v_h}\int dk_e \int dk_h\ \varphi_\sigma^{v_e v_h}(k_e k_h)\, e^{i(k_e\rho_e - k_e\rho_e)}\, \xi_{v_e k_e}^{j_e}(z_e)\, \xi_{v_h k_h}^{j_h}(z_h)\,, \qquad (36)$$

where $j_{e,h}$ indicates the spinor components of a multi-component electron or hole subband function, $k_{e,h}\rho_{e,h}$ are in-plane wave vectors (space vectors) and $\varphi_\sigma^{v_e v_h}(k_e k_h)$ describes the mixture of different pairs of electron and hole subband states into a quantum well exciton state with internal quantum numbers σ. The mixing is caused by the *e-h* Coulomb interaction. The excitonic absorption, as calculated following Fermi's Golden Rule, has the same structure as shown in (22), but now the excitation is from the electronic ground state to correlated excited (or exciton) states and the matrix element between these states has to be taken from the many-particle dipole operator. It is convenient to use the excitonic oscillator strength per unit area A

$$f_x^\sigma = \frac{2}{m_0 A \mathcal{E}_\sigma}\left|\langle 0|e\bullet\sum_i \mathbf{p}_i|\psi_\sigma\rangle\right|^2 \qquad (37)$$

and to write

$$\alpha_x(\omega) \sim \sum_\sigma f_x^\sigma\,\delta(\mathcal{E}-\mathcal{E}_\sigma) + \int d\mathcal{E}'\, f_x^\sigma(\mathcal{E}')D(\mathcal{E}')\delta(\hbar\omega-\mathcal{E}')\,. \qquad (38)$$

$\alpha_x(\omega)$ consists of a discrete and a continuous part, corresponding to bound and scattering states of the relative motion.

Among the many contributions to the theory of excitons in quantum wells, we mention only more recent ones (Pasquarello and Andreani, 1990; Chao and Chuang, 1993; Winkler, 1994), which stress the aspect of mixing of electron-hole pairs from different subbands by the Coulomb interaction. As a demonstration of this effect and of the capability of the concepts to provide quantitative explanation of experimental data we show in Fig. 9 data for a multiple quantum well system with 22.5 nm wide GaAs wells and 12.0 nm wide $Al_{0.35}Ga_{0.65}As$ barriers from photoluminescence (a) and calculations for a single such quantum well with a decomposition into contributions from different *e-h* subband pairs (Winkler, 1994). Besides a close agreement between the experimental and calculated spectra in the position and intensity of the structures, the decomposition gives evidence, that the simple assignment indicated in the experimental data is not justified.

CONFINED ELECTRONS AND HOLES

The quantum structures discussed so far in my lectures do not contain free carriers, which are essential for transport and will be the subject of this section. In bulk semiconductors free carriers can be introduced by doping: replacing an atom of the host lattice by an impurity atom with more (or less) electrons eventually creates impurity states close to the conduction (valence) band edge. The carriers in these states can easily be excited (e.g.~by increasing the temperature) to the conduction (valence) band continuum where they are available as free electrons (holes); therefore, we talk about donors (acceptors). Obviously there are two aspects of doping: providing free carriers and introducing impurities. As was emphasized in the first section the carrier mean free path at

low temperatures is limited by elastic scattering from impurities. Thus, in bulk material we can increase the carrier density only at the prize of lowering the mobility (at low temperature). In semiconductor structures it is possible to spacially separate the free carriers from the ionized impurities and thus increase the mobility by many orders of magnitude. These high mobility 2D electron and hole gases are basic for observation of mesoscopic phenomena. Their (single particle and collective) excitation spectra provide fundamental insight into correlation effects of these many-particle systems. 1D and 0D semiconductor structures or lateral superlattices for investigation of transport or correlation are always obtained by imposing lateral structure onto a high-mobility 2DES.

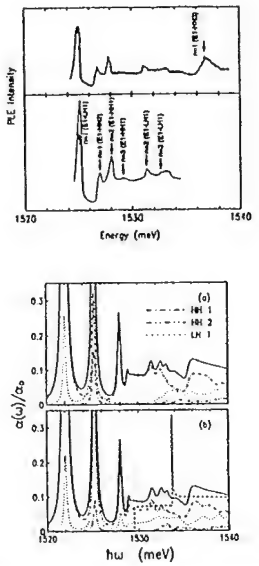


Figure 9. PLE spectra and assignment of transitions (upper part) and calculated excitonic absorption spectrum with decomposition into contributions from different hole subbands (lower part) (from Winkler, 1994).

Inversion Layers, Modulation Doping

The first experimental demonstration of a 2D electron system was performed already 1966 for an inversion layer in a MOS device (Fowler *et al.*, 1966). In this experiment SdH oscillations at 1.34 K for a fixed magnetic field perpendicular to the inversion layer have been measured by varying the gate voltage V. The SdH period turned out independent of V, indicating the same number of states in each Landau level (see (21)) as characteristic for a 2D system. Because of the progress in planar Si technology, 2D electron and hole systems in inversion layers of Si-MOS-devices have been the main subject also of fundamental research (Ando et al., 1982) up to the discovery of the quantum Hall effect by K. von Klitzing (1990; von Klitzing *et al.*, 1980).

Two material specific aspects are responsible for a limitation of electron mobilities in Si-MOS: the interface between Si and SiO_2 is not planar on an atomic scale and the electron mass is too large. Here semiconductor structures based on epitaxial growth (MBE, MOCVD) of III-V semiconductor structures have the advantage of more perfect interfaces and smaller electron masses. In the meantime even Si technology makes use of MBE in Si/Ge heterostructures and strained layer systems. The essential concept to realize extremely high mobilities is modulation or remote deping (Dingle *et al.*, 1978): by placing the impurity atoms in the large gap material, the carriers --- which are collected at the interface but in the small gap material --- are spacially separated from the scatterers (Fig. 10). This effect can be increased by introducing a spacer and has led to record mobilities in AlGaAs/GaAs systems which exceed 10^7 cm^2/Vs and correspond to elastic mean free paths of 10 μm (Foxon et al., 1989).

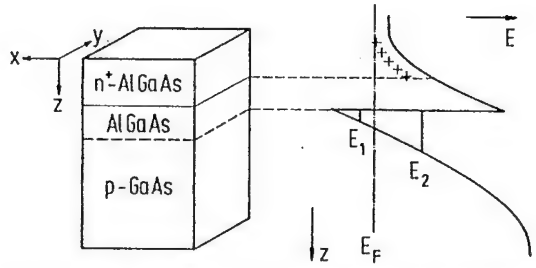


Figure 10. Real space picture (left) and energy diagram for a heterostructure with remote doping (right).

With the appearance of doping and free carriers in separate parts of the semiconductor structure the particles experience a potential, which has to be determined by solving self-consistently the single-particle Schrödinger equation, e.g.~with Hamiltonian of (30) for electrons or of (34) for holes, but now the single-particle potential contains besides the z-dependent band edge energies Hartree and exchange-correlation potentials in the vein of density-functional theory (DFT), see Ando *et al.* (1982). The Hartree potential is obtained by solving Poisson's equation

$$\frac{\partial^2}{\partial z^2} V_H(z) = \frac{e}{\varepsilon\varepsilon_0}\rho(z) \quad , \tag{39}$$

for a given charge distribution $\rho(z)$, which consists of charged impurities and occupied subband states. Charge neutrality has to be observed as boundary condition. The exchange correlation potential V_{XC} is usually adopted as an analytic expression using the local density approximation (LDA) and some elaborate numerical results for the homogeneous electron system (Ando *et al.*, 1982). Differences in the dielectric constants across the interface give rise to an image potential $V_{im}(z)$.

As for undoped systems, the spectrum of confined electrons and holes in doped heterostructures and quantum wells shows signatures of the underlying bulk band structure, which can be considered on different levels of accuracy (Winkler and Rössler, 1994). In contrast to the undoped quantum structures, probing of the doped structures becomes more complicated because the system responds to the external perturbation as a many-particle system. Thus depending on the probing of the system, the spectral features are eventually

not of single-particle nature but exhibit collective effects. This will be the case irrespective of the dimensionality of the system.

Single Particle and Collective Excitations

The standard experimental setup for FIR transmission is connected with an electric field vector in the plane of the 2DES, which cannot excite an intersubband transition, but a cyclotron resonance if a static magnetic field component in z-direction is present. The static magnetic field **B** can be used to tune the cyclotron energy $\hbar\omega_c$ to the intersubband resonance. A small tilt of **B** out of the z-direction couples the two transitions and allows to probe the intersubband separation by measuring cyclotron resonance. The response function for this experiment is the dynamical conductivity (Ando *et al.*, 1982). It describes the response of the many-body system of electrons (or holes) in a doped quantum structure to the external perturbation, $V_{ext} = eFz\ e^{i\omega t + iq\cdot r}$ (the scalar potential connected with the z-component of the FIR electromagnetic field), and to the changes of the Hartree and exchange-correlation potential due to the induced charge density $\Delta n(z)$. Thus the system responds linearly to the total perturbation

$$H'e^{i\omega t} = [eFz + \Delta V_H(z) + \Delta V_{XC}(z)]\ e^{i\omega t + iq\cdot r}\ , \tag{40}$$

where

$$\Delta V_H(z) = -\frac{4\pi e^2}{\varepsilon\varepsilon_0}\int_0^z dz' \int_0^{z'} dz''\ \delta n(z'') \tag{41}$$

and

$$\Delta V_{XC}(z) = \frac{\partial V_{XC}}{\partial n(z)}\ . \tag{42}$$

In terms of subband functions $\xi_\nu(z)$ and excitation energies out of the lowest subband (the single one assumed to be occupied, i.e.~electrical quantum limit) we find by time-dependent perturbation theory the induced charge density

$$\Delta n(z) = -\sum_{\nu\nu'} \Pi_{\nu\nu'}(q,\omega)\xi_\nu(z)\xi_{\nu'}(z)\int_{-\infty}^{\infty} dz'\ \xi_\nu(z')\xi_{\nu'}(z')H'\ , \tag{43}$$

where

$$\Pi_{\nu\nu'}(q,\omega) = \sum_k f_0[\mathcal{E}_\nu(k)]\ 2\frac{\mathcal{E}_{\nu'}(k+q) - \mathcal{E}_\nu(k)}{[\mathcal{E}_{\nu'}(k+q) - \mathcal{E}_\nu(k)]^2 - [\hbar\omega]^2}\ . \tag{44}$$

Linear response in the electrical quantum limit means to replace according to

$$\langle 1,k|V_{ext}|\nu,k+q\rangle = \sum_{\nu'} \langle 1,k|H'|\nu',k+q\rangle M_{\nu\nu'}(q,\omega)\ , \tag{45}$$

where the matrix $M(q,\omega)$ with elements

$$M_{\nu\nu'}(q,\omega) = \delta_{\nu\nu'} - \Pi_{\nu'}(q,\omega)[\alpha_{\nu\nu'}(q) + \beta_{\nu\nu'}] \tag{46}$$

describes the response of the 2DES to an external perturbation with wave vector **q** and frequency ω. The collective modes of the system are derived from

$$\det[M(q,\omega)] = 0\ . \tag{47}$$

For FIR excitation $q \sim 0$; we find for the special case of excitation from the lowest occupied to the second subband the excitation energy

$$\hbar\tilde{\omega}_{21} = (\mathcal{E}_2 - \mathcal{E}_1)[1 + \alpha_{11} - \beta_{11}]^{1/2}\ , \tag{48}$$

where

$$\alpha_{11} = 2N_s\frac{4\pi e^2}{\varepsilon\varepsilon_0}\frac{1}{\mathcal{E}_2 - \mathcal{E}_1}\int_{-\infty}^{\infty} dz\ [\xi_2'(z)\xi_1(z) - \xi_1'(z)\xi_2(z)] \tag{49a}$$

and

$$\beta_{11} = -2N_s\frac{1}{\mathcal{E}_2 - \mathcal{E}_1}\int_{-\infty}^{\infty} dz\ \xi_2^2(z)\xi_1^2(z)\frac{\partial V_{XC}(z)}{\partial n(z)}\ . \tag{49b}$$

These corrections, which have their origin in the induced changes of the Hartree and exchange-correlation potential, respectively, are called depolarization shift and excitonic correction. The response function for the transmission experiment is the absorbed power per unit area or else $\mathrm{Re}\{\sigma_{zz}\omega\}$, which reads:

$$\sigma_{zz}\omega = -i\omega\frac{N_s e^2}{m^*}\frac{f_{21}}{\tilde{\omega}_{21}^2 - \omega^2 - 2i\omega/\tau}\ . \tag{50}$$

This expression for the dynamical conductivity shows that for doped quantum structures the excitation does not occur at the single particle energy, it is shifted due to many-body corrections. These corrections depend on the particle density (Wieck *et al.*, 1989; Ensslin *et al.*, 1989).

Besides FIR spectroscopy in particular optical spectroscopy with transitions between electron and hole subbands have been used to detect the many-particle aspects of 2DES. Inelastic light scattering in degenerate bulk semiconductors is known to disclose single particle and many-body excitations by playing with the polarization of incoming and scattered light (Abstreiter *et al.*, 1988). This concept has been successfully applied more recently (Pinczuk, 1992) to detect and distinguish single particle excitations from spin-density and charge density excitations. Polarization dependent luminescence experiments have been used as well to detect many particle effects in spin-relaxation processes (Uenoyama and Sham, 1990).

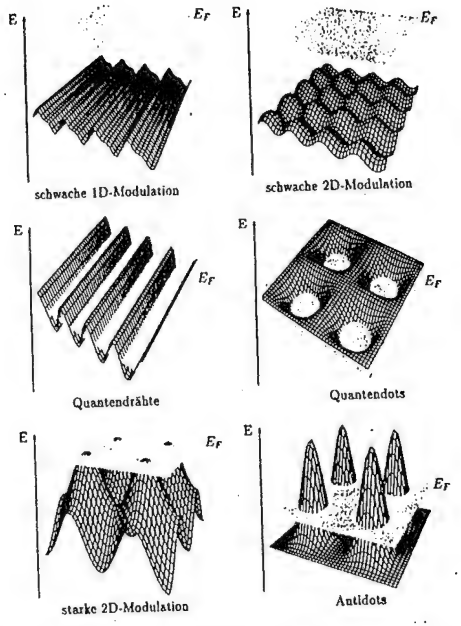


Figure 11. Schematic picture of lateral periodic potentials. The Fermi energy ε_F in comparison with the strength of the modulation allows to distinguish between weak and strong modulation.

WIRES AND DOTS

Various concepts have been developed and are presently used to impose lateral structures to a 2DES (see lectures of M. Reed in this volume). Single dots or wires are usually realized with split-gate techniques which allow to deplete the carriers under the gate by tuning the gate voltage. Static transport through isolated point contacts turns out to be quantized even without a magnetic field (van Wees *et al.*, 1988). Isolated dots with tunable point contacts to the leads have been used to demonstrate charging effects and the transport of single electrons (Meirav *et al.*, 1991). Periodic arrays of wires and dots are realized by lithographic techniques in combination with etching or adding of metallic gates. The latter structures are designed to change the strength of the electrostatic potential modulation in the plane of the 2DES and allows to study an increasing influence of the lateral structure (Fig. 11). Frequently transmission through such lateral structures are investigated in the presence of a magnetic field: for sufficiently strong magnetic fields the magnetic length l_B gets smaller than the period of the lateral structure and dimensional crossover can be observed. These structures are investigated by FIR spectroscopy, described by the dynamical conductivity, and reveal informations on the number of electrons and the lateral confinement potential (Hansen *et al.*, 1992). As a special system with nanostructures antidot superlattices have recently attracted attention. In these systems the lateral potential modulation is so strong, that it is seen by the carriers as strongly repulsive barriers (antidots) (Weiss *et al.*, 1991). This system is a special version of a billiard, which in the concepts of classical nonlinear dynamics shows chaos (Fleischmann et al., 1992).

Quantum Dots: Artificial Atoms

Lateral structuring of a 2DES opens a way to create systems with a small number of electrons (or holes), which are electrostatically confined. A change of the confinement potential, e.g.~by tuning a gate voltage, eventually changes the number of electrons in the system. These few electron systems are sometimes addressed as artificial atoms (Kastner, 1992, 1993): their characteristic length (diameter of a lateral confinement potential L, or oscillator length of a parabolic confinement potential $(\hbar/m\omega_0)^{1/2}$ is of the order of nm, the characteristic confinement energy of the order *meV*. Remember, in natural atoms we have instead the Bohr radius and the Rydberg constant, respectively, both quantities directly related to the Coulomb interaction. In artificial atoms, we can change length and energy independently by changing the mass or the confinement.

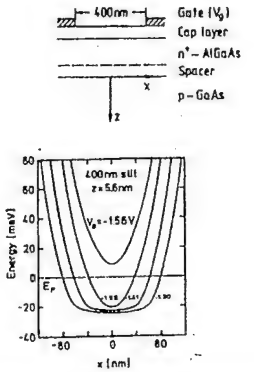


Figure 12. Gate geometry and selfconsistent potential for a quantum wire (after Laux *et al.*, 1988; Kumar *et al.*, 1990).

First evidence of a zero-dimensional semiconductor nanostructure (or quantum dot) came from a transport experiment through an etched double barrier structure (Reed *et al.*, 1989), as mention already in the first section. Self-consistent calculations of gated 1D and 0D structures have been performed by solving the Schrödinger and Poisson equations with boundary conditions defined by the gate geometry (Laux *et al.*, 1988; Kumar *et al.*, 1990). These calculations show a parabolic confinement potential --- which seems to be characteristic for any electrostatic confinement --- which flattens out due to the Hartree potential if more and more electrons fill the structure (Fig. 12).

With these results in mind, a simple model becomes possible for electrons in quantum dots: the two-dimensional harmonic oscillator, which even in the presence of a perpendicular magnetic field can be solved exactly (Fock, 1928). The Hamiltonian

$$H = \frac{1}{2m}(\mathbf{p} + e\mathbf{A})^2 + \frac{1}{2}m\omega_0^2(x^2 + y^2) \ , \qquad (51)$$

with $\mathbf{A} = (B/2)(-y,x,0)$ in the symmetric gange can be cast into the form (Rössler *et al.*, 1992)

$$H = \hbar\omega_+(a_+^+a_+ + \frac{1}{2}) + \hbar\omega_-(a_-^+a_- + \frac{1}{2}) \quad , \tag{52}$$

by introducing right and left hand oscillator operators $a_\pm$, $a_\pm^+$. The frequencies

$$\omega_\pm = \{\omega_0^2 + \frac{1}{2}\omega_c^2\}^{1/2} \tag{53}$$

expose the two principle dipole excitations of the system (for right and left hand circularly polarized light). The energy spectrum as a function of the magnetic field is shown in Fig. 13. The ω_+ mode corresponds to a closed classical orbit, it approaches the cyclotron frequency ω_c with increasing magnetic field. The ω_- mode can be ascribed to a classical orbit shipping along the confinement potential (in another context it is called an edge-plasmon).

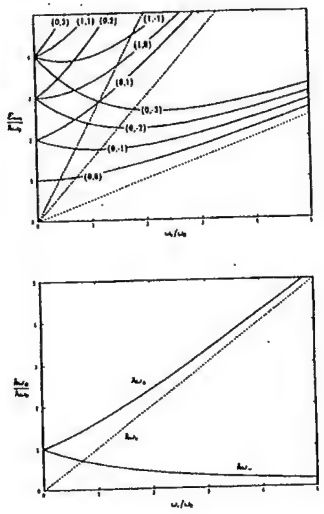


Figure 13. Energy eigenvalues of an electron in a parabolic quantum dot (upper part) and dipole excitation energies $\omega_\pm$ (lower part) vs. magnetic field.

This two mode behavior is characteristic for parabolic confinement. It will even persist in dots with N electrons due to a generalization to Kohn's theorem. A dot with N electrons would be described by

$$H = \sum_{i=1}^{N} \left(H_i + \frac{1}{2}\sum_{j\neq i}\frac{e^2}{\varepsilon|\mathbf{r}_i - \mathbf{r}_j|}\right) , \tag{54}$$

with H_i being the single particle Hamiltonian of (51) or (52) for the i-th electron. The dipole operator

$$H_{dip} = \sum_{i=1}^{N} e\mathbf{E}\bullet\mathbf{r}_i = Ne\,\mathbf{E}\bullet\mathbf{r} \tag{55}$$

couples to the center of mass motion --- which is again of the type (52) --- thus causing transitions only at frequencies $\omega_\pm$. Deviations from parabolic confinement as well as from a parabolic band structure can cause changes of this spectrum, which partially depend on a violation of Kohn's theorem. This has been studied in detail for dots with two electrons (Gudmundsson and Gerhardts, 1991; Pfannkuche and Gerhardts, 1991; Darnhofer and Rössler, 1993; Junker et al., 1994).

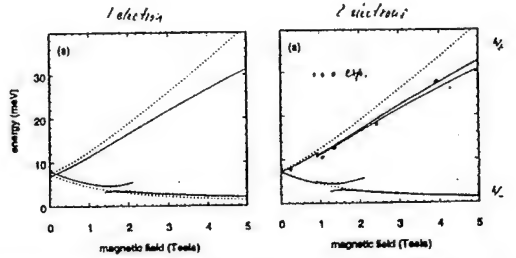


Figure 14. Dipole excitation energies vs. magnetic field for 1 and 2 electrons in a parabolic quantum dot but nonparabolic bandstructure (parameter of InSb). The dotted lines correspond to a parabolic bandstructure. Experimental data are given for comparison (after Rössler *et al.*, 1992; Gudmundsson and Gerhardts, 1991; Pfannkuche and Gerhardts, 1991; Darnhofer and Rössler, 1993; Junker *et al.*, 1994).

The energy spectrum of dots with one and two electrons is shown in Fig. 14 for system parameters typical for InSb including nonparabolicity and spin-orbit coupling. From these spectra we obtain the dipole excitation energies, which can be compared with experimental data. Except for small deviations the calculated dipole excitation energies are almost identical for the 1 and 2 electron system, which essentially shows the two mode behavior: the ω_+ frequency is reduced due to nonparabolicity, the ω_- mode shows a level anticrossing due to spin-orbit coupling. In addition in the 2 electron data we find a small splitting of the ω_+ mode connected with a change of the ground state spin configuration at about 1 Tesla and coupling between relative and center-of-mass motion, i.e.~violation of Kohn's theorem.

While the calculations for quantum dot helium $N = 2$ are simple enough to consider deviations from parabolic confinement or band structure, they become rather complex with increasing N even without these deviations. Results for $N < 10$ are available from different concepts. Direct diagonalization of the Hamiltonian (54) in a basis of antisymmetrized products of single particle states has been accomplished up to $N = 5$ (Hawrylak and Pfannkuche, 1993, and references contained therein). Results from Quantum Monte-Carlo calculations are available for $N < 10$ (Bolton, 1994). The characteristic feature is a change of the ground state spin configuration with the magnetic field as shown in Fig. 15 for N =3,4. Experimental evidence for this feature comes from capacitance spectroscopy on single dots.

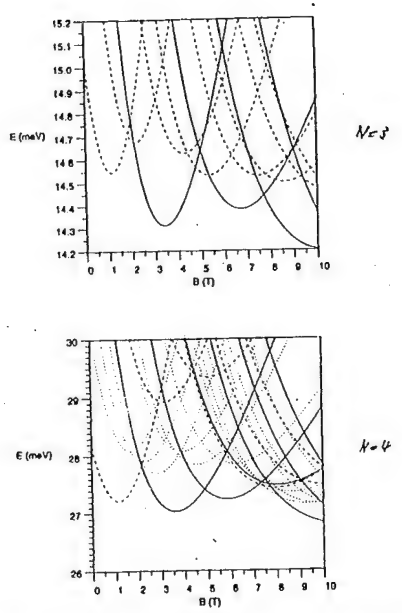


Figure 15. Groundstate energies of $N = 3$ and 4 electrons in a parabolic quantum dot vs. magnetic field from quantum Monte-Carlo calculations (from Bolton, 1994).

Transport in Lateral Superlattices

A lateral potential modulation of the 2DES leads to significant changes of the magnetotransport properties. In Fig. 16 experimental data (Weiss et al., 1991) of the longitudinal and Hall resistivity are compared for 2D electron systems without and with lateral superlattice as indicated in the inset of Fig. 16. Without modulation R_{xx} shows the typical Shubnikov-de Haas oscillations and R_{xy} the monotonic increase with B and the quantum Hall plateaus. These features are slightly changed by the modulation, but in R_{xx} additional strong peaks appear at low magnetic fields and the Hall effect is quenched. These modifications, in particular the additional peaks in R_{xx}, have been ascribed to pinning of electrons on commensurate orbits around 1, 2, 4, 9,... antidots. Classical nonlinear dynamics studies give support to this interpretation but point out as well the importance of chaotic aspects in this system (Fleischmann *et al.*, 1992).

The system Hamiltonian is

$$H = \frac{1}{2m}(\mathbf{p} + e\mathbf{A})^2 + V(x,y) \tag{56}$$

with a 2D periodic potential (lattice constant a)

$$V(x,y) = V_0\left[\cos\left(\frac{\pi x}{a}\right) + \cos\left(\frac{\pi y}{a}\right)\right]^{2/\alpha} . \tag{57}$$

Solutions of the classical equations of motion for an electron with given energy reveal the strong dependence on the initial condition (position and velocity) of the motion. Depending on the strength of the magnetic field the Poincaré plots show the mixed phase space with a tendency from chaotic to more regular behavior with increasing B (s.~Fig. 17). With increasing magnetic field l_B gets smaller and the electrons can perform regular cyclotron orbits between antidots instead of being scattered at the dots (competition between a and l_B!).

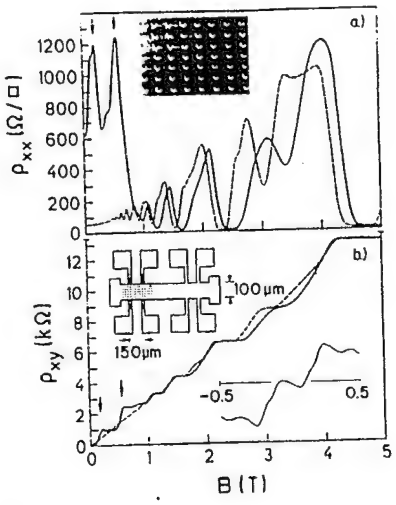


Figure 16. Longitudinal ρ_{xx} and Hall resistance ρ_{xy} for a 2DES without (dashed lines) and with lateral periodic potential (solid lines) (from Weiss *et al.*, 1991).

Although the system parameters (a = 300 nm and $\lambda_F \sim 60$ nm) in the experiments of Weiss *et al.* (1991) lend support to consider the situation as classical, a theoretical study using quantum mechanical concepts seems to be in place in view of future trends to smaller lattice constants. As a first step one has to find the quantum mechanical eigensolutions for the Hamiltonian (56). This has been achieved by expanding the solutions in a series of magnetic Bloch functions (Silberbauer, 1992). This concept is restricted to so-called rational magnetic field values $B = r\,\Phi_0/a^2$ with r an integer and the magnetic flux quantum $\Phi_0 = h/e$, for which the eigenstates $|n,\theta>$ of H (56) can be classified by a miniband index n and a magnetic wave vector θ. In Fig. 18, results for two different magnetic field values an θ along two symmetry lines of the magnetic Brillouin zone are shown together with the potential and a plane indicating the Fermi energy. it is important to see that the quite irregular minibands at low magnetic fields evolve into almost dispersionless ones at higher magnetic fields, which tend to cluster towards the Landau levels of the unmodulated 2DES.

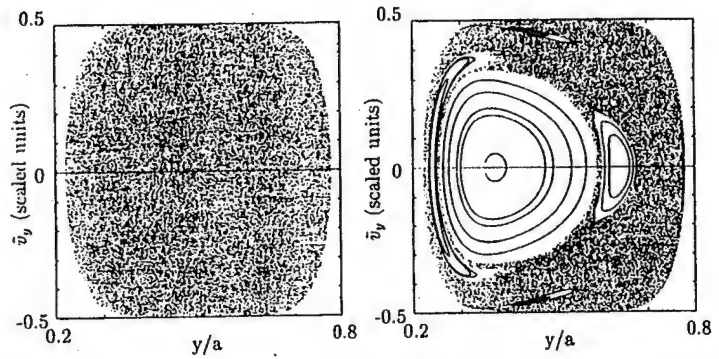


Figure 17. Poincare plots for classical motion of electrons described by the Hamiltonian of (56) for two different values of the magnetic field (from Silberbauer *et al.*, 1994).

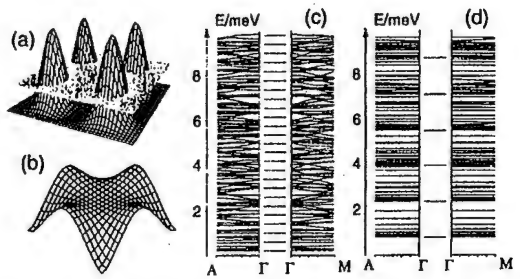


Figure 18. Antidot potential (a), miniband (b) and dispersion of minibands along main symmetry lines for B = 0.23 T (c) and 0.92 T (d) (from Silberbauer *et al.*, 1994).

The transport properties depend on one hand on the solutions of the quantum mechanical problem but are determined essentially by the scattering of the carriers by impurities. This has to be considered in evaluating the Kubo formulas for the conductivity tensor which in units of e^2/h can be written

$$\sigma_{\mu\mu} = \frac{2\pi}{r}\hbar\omega_c^2 \int d\mathcal{E}\ \left(-\frac{df}{d\mathcal{E}}\right) \sigma_{\mu\mu}(\mathcal{E}) \quad , \tag{58}$$

with

$$\sigma_{\mu\mu}(\mathcal{E}) = \sum_{n_1 n_2} \int d^2\theta\ |< n_1\theta|k_\mu|n_2\theta>|^2\ A_{n_1\theta}(\mathcal{E})\ A_{n_2\theta}(\mathcal{E}) \tag{59}$$

for the longitudinal conductivity and

$$\sigma_{\mu\nu} = \frac{4}{r}\hbar\omega_c^2 \int d\mathcal{E}\ f(\mathcal{E})\ \sigma_{\mu\nu}(\mathcal{E}) \quad , \tag{60}$$

with

$$\sigma_{\mu\nu}(\mathcal{E}) = \sum_{n_1 \neq n_2} \int d^2\theta\ (< n_1\theta|k_\mu|n_2\theta>< n_2\theta|k_\nu|n_1\theta>) \frac{dG_{n_1\theta}}{d\mathcal{E}} A_{n_2\theta}(\mathcal{E}) \tag{61}$$

for the Hall conductivity. $G_{n\theta}(\mathcal{E})$ and $A_{n\theta}(\mathcal{E})$ are the Green's function and spectral function, respectively. Impurity scattering is considered in the self-consistent Born approximation (SCBA) by solving the self-consistency equation

$$\Sigma(z) = \gamma^2\ \mathrm{tr}\{G(z)\} \tag{62}$$

for the self-energy. For simplicity the self-energy is assumed to be independent of n and θ and $G(z)$ is taken as impurity averaged Green's function. The parameter γ can be traced back to the mobility of the unmodulated 2DES and thus be determined from experimental data.

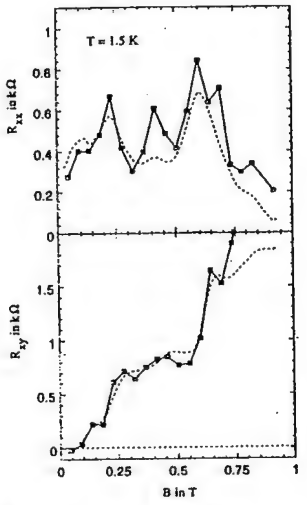


Figure 19. Calculated (symbols and solid line) and experimental (dashed lines, from Weiss *et al.*, 1991) magneto resistance for a 2DES antidot superlattice (from Silberbauer *et al.*, 1994).

These calculations (Silberbauer and Rössler, 1994) have been performed on an absolute scale (see Fig. 19) and yield a surprisingly good agreement with the experimental data of Weiss *et al.* (1991) (dashed lines). All features, the commensurability peaks in R_{xx}, the disturbed Hall plateaus in R_{xy} and the quenched Hall effect at low magnetic field are reproduced. This agreement gives strong support to the quantum mechanical concept (including the approximations involved in evaluating the impurity scattering) of magneto-transport calculations. As the magnetotransport data can be interpreted either in terms of classical nonlinear dynamics (including chaos) or in a quantum mechanical picture one may

ask now for the fingerprints of chaos in the quantum mechanical model. This has been done recently (Silberbauer et al., 1994) by applying level statistics to the eigenvalues of (56).

REFERENCES

Abrikosov, A. A., Gorkov, L. P., and Dzyaloshinskii, I. Ye., 1965, "Quantum Field Theoretical Methods in Statistical Physics," Pergamon Press, Oxford.

Abstreiter, G., Merlin, R., and Pinczuk, A., 1988, *IEEE J. Quantum Electron.* QE-22.

Anderson, P. W., Abrahams, E., and Ramakrishnan, T. V., 1979, *Phys. Rev. Lett.* 43:718.

Ando, T., Fowler, A. B., and Stern, F., 1982, *Rev. Mod. Phys.* 54:437.

Andreani, L. C., Pasquarello, A., and Bassani, F., 1987, *Phys. Rev. B* 36:5887.

Ashcroft, N. W., and Mermin, N. D., 1976, "Solid State Physics," Holt, Rinehart and Winston, New York.

Bastard, G., Brum, J. A., and Ferreira, R., 1991, *Solid State Physics* 44:229.

Bate, R. T., 1988, *Sci. Am.* 258:78.

Beaumont, S. P., and Sotomayer-Torres, C. M., 1990, "Science and Engineering of one-and zero-dimensional semiconductors," Plenum Press, New York, NATO ASI, Physics Series B214.

Beenakker, C. W. J., and van Houten, H., 1991, *in*: " Solid State Physics," H. Ehrenreich, ed., Academic Press, New York, Vol. 44:1.

Bergmann, B., 1983, *Phys. Rev. B* 28:2914.

Bloch, F., 1928, *Z. Phys.* 52:555.

Bolton, F., 1994, *Sol.-State Electron.* 37:1159.

Büttiker, M., 1986, *Phys. Rev. Lett.* 57:1761.

Capasso, F., 1987, *in*: "Physics and Applications of Quantum Wells and Superlattices," E.E. Mendez, and K. von Klitzing, eds., Plenum Press, New York, NATO ASI Series B170:377.

Chang, L. L., Esaki, L., and Tsu, R., 1974, *Appl. Phys. Lett.* 24:593.

Chao, C. Y.-P., and Chuang, S. L., 1993, *Phys. Rev. B* 48:8210.

Darnhofer, T., and Rössler, U., 1993, *Phys. Rev. B* 47:16020.

Dawson, P., Moore, K. J., Duggan, G., Ralph, H. I., and Foxon, C. T. B., 1986, *Phys. Rev. B* 34:6007.

Dingle, R., 1975, *in*: "Festkörperprobleme/Adv. in Solid State Physics XV," H.J. Queisser, ed., Pergamon-Vieweg, Braunschweig, 21.

Dingle, R., Wiegmann, W., and Henry, C. H., 1974, *Phys. Rev. Lett.* 33:827.

Dingle, R., Stoermer, H. L., Gossard, A. C., and Wiegmann, W., 1978, *Appl. Phys. Lett.* 33:665.

Duchenin, J. P., *et al.*, 1985, *in*: "MBE and Heterostructures," L. L. Chang and K. Ploog, eds., Nijkoff Publishers, Dordrecht, NATO ASI, Series E87:677.

Ensslin, K., Heitmann, D., and Ploog, K., 1989, *Phys. Rev. B* 39:10879.

Esaki, L., 1991, *in*: "Electronic properties of multilayers and low-dimensional semiconductor structures," J.M. Chamberlain, L. Eaves, J.-C. Portal, eds., Plenum Press ," Plenum Press, New York, NATO ASI, Physics Series B 231:1.

Esaki, L., and Tsu, R., 1970, IBM J. Res. Dev. 14:61.

Fasol, G., Fasolino, A., and Lugli, P., 1990, "Spectroscopy of semiconductor microstructures," Plenum Press, New York, NATO ASI B206.

Feldmann, J., 1992, *in*: "Festkörperprobleme/Adv. in Solid State Physics 32," U. Rössler, ed., Vieweg, Braunschweig.

Ferry, D. K., and Grubin, H. L., 1994, *in*: "Solid-State Physics," H. Ehrenreich, Ed., Academic Press, New York.

Fock, V., 1928, *Z. Phys.* 47:446.

Fowler, A. B., Fang, F. F., Howard, W. E., and Stiles, P. J., 1966, *Phys. Rev. Lett.* 16:901; *J. Phys. Soc. Japan* 21(Suppl.):331.

Fleischmann, R., Geisel, T., and Uetzmenck, R., 1992, *Phys. Rev. Lett.* 68:1367.

Foxon, C. T., Harris, J. J., Hilton, D., Hewett, J., and Roberts, C., 1989, *Semicond. Sci. Technol.* 4:582.

Gudmundsson, V., and Gerhardts, R., 1991, *Phys. Rev. B* 43:12098.

Hansen, W., Kotthaus, J. P., and Merkt, U., 1992, *in*: "Semicond. & Semimetals," Vol. 35, M. Reed, ed., Academic Press, San Diego.

Hayden, R. K., *et al.*, 1991, *Phys. Rev. Lett.* 66:1749.

Heiblum, M., *et al.*, 1987, *Phys. Rev. Lett.* 58:816.

Goldoni, G., and Fasolino, A., 1992, *Phys. Rev. Lett.* 69:2567.

Hawrylak, P., and Pfannkuche, D., 1993, *Phys. Rev. Lett.* 70:485.

Junker, P., Kops, U., Merkt, U., Darnhofer, T., and Rössler, U., 1994, *Phys. Rev. B* 49:4794.

Kajikawa, Y., 1993, *Phys. Rev. B* 48:7935.

Kastner, M. A., 1992, *Rev. Mod. Phys.* 64:849.

Kastner, M. A., 1993, *Physics Today*, Jan., p. 24.

Kumar, A., Laux, S. E., and Stern, F., 1990, *Phys. Rev. B* 42:5166.

Landauer, R., 1957, *IBM J. Res. Develop.* 1:223.

Landauer, R., 1988, *IBM J. Res. Develop.* 32:306.

Landauer, R., 1989, *Phys. Today* 42(October):119.

Laux, S. E., Frank, D. J., and Stern, F., 1988, *Surf. Sci.* 196:101.

Meirav, U., *et al.*, 1991, *Z. Phys. B -- Cond. Matter* 85:357.

Mendez, E. E., 1993, *Physics Today* .

Mendez, E. E., Wang, W. I., Ricco, B., and Esaki, L., 1977, *Appl. Phys. Lett.* 47:415.

Narayanamurti,, V., 1984, Physics Today October, 1984:24.

Pasquarello, A., and Andreani, L. C., 1990, *Phys. Rev. B* 42:8928.

Pfannkuche, D., and Gerhardts, R., 1991, *Phys. Rev. B* 44:13132.

Pinczuk, A., 1992, *in*: "Festkörperprobleme/Adv. in Solid State Physics 32," U. Rössler, ed., Vieweg, Braunschweig, 45.

Reed, M. A., *et al.*, 1989, *in*: "Festkörperprobleme/Adv. in Solid State Physics 29," U. Rössler, ed., Vieweg, Braunschweig, 267.

Reynolds, D. C., and Collins, T. C., 1981, *in*: "Excitons, their properties and uses," Academic Press, New York.

Rössler, U., 1979, *in*: "Festkörperprobleme/Adv. in Solid State Physics XIX," J. Treusch, ed., Vieweg, Braunschweig, 77.

Rössler, U., Broido, D. A., and Bolton, F., 1992, *in*: "Low-dimensional Elelctronic Systems," G. Bauer, F. Kuchar, H. Heinrich, eds., Springer, Berlin, 21.

Rosencher, E., and Vinter, B., 1992, "Intersubband transitions in Quantum Wells," Plenum Press, New York, NATO ASI, Physics Series B288.

Silberbauer, H., 1992, *J. Phys. C* 4:7355.

Silberbauer, H., and Rössler, U., 1994, submitted to *Phys. Rev. Lett.*

Silberbauer, H., Rotter, P., Suhrke, M., and Rössler, U., 1994, Proc. Int. Winterschool, Mauterndorf, H. Heinrich, G. Bauer, and F. Kuchar, eds., in press.

Sollner, T. C. L. G., Brown, E. R., Parker, C. D., and W.D. Goodhue, W. D., 1991, *in*: "Electronic properties of multilayers and low-dimensional semiconductor structures," J. M. Chamberlain, L. Eaves, J.-C. Portal, eds., Plenum Press, New York, NATO ASI B231:1.

Tsao, J. Y., 1993, *in*: "Materials Fundamentals of Molecular Beam Epitaxy," Academic Press, New York.

Tsu, R., and L. Esaki, L., 1973, *Appl. Phys. Lett.* 22:562.

Uenoyama, T., and Sham, L. J., 1990, *Phys. Rev. B* 42:7114.

van Wees, B. J., *et al.*, 1988, *Phys. Rev. Lett.* 60:848.

von Klitzing, K., 1990, *in*: "Festkörperprobleme/Adv. in Solid State Physics 30," U. Rössler, ed., Vieweg, Braunschweig, 25.

von Klitzing, K., Dorda, G., and Pepper, M., 1980, *Phys. Rev. Lett.* 45:494.

Waschke, C., Roskos, H. G., Schwedler, R., Leo, K., Kurz, H., and Köhler, K., 1993, *Phys. Rev. Lett.* 70:3319.

Weiss, D., Roukes, M. L., Menschig, A., Grambow, P., von Klitzing, K., and Weimann, G., 1991, *Phys. Rev. Lett.* 66:2790.
Wieck, A. D., *et al.*, 1989, *Phys. Rev. B* 39:3785.
Winkler, R., 1994, PhD Thesis, Regensburg (to be published).
Winkler, R., and Rössler, U., 1993, *Phys. Rev. B* 48:8918.
Winkler, R., and Rössler, U., 1994, *Surf. Sci.* 305:295.

FABRICATION OF NANOSCALE DEVICES

Mark A. Reed and Jeffrey W. Sleight
Departments of Electrical Engineering and Applied Physics
Yale University
New Haven, CT 06520

INTRODUCTION

The toolbox of nanofabrication technologies that has developed as a result of microelectronic needs has given researchers an unprecedented capability to design and construct quantum effect and mesoscopic device structures. The subsequent turning point in the understanding of nanometer scale electronic transport was the development of reliable fabrication techniques on the nanometer scale. Laboratories around the world can now controllable impose additional lateral dimensions of quantum confinement on 2D systems with length scales approaching those of epitaxial lengths in the growth direction. The achievement of quantum wires, dots, and Coulomb blockade structures presented near-to ultimate limit electronic systems to the experimentalist.

How does one fabricate these structures of low dimensionality? The obvious approach is to utilize the existing technology of MBE or MOCVD to define a 2D system, and impose additional lateral confinement with nanometer lithographic techniques. State-of-the-art in electron-beam lithography can define dimensions in the 10nm regime, with pattern transfer techniques in the same dimensional regime, clearly sufficient to observe large quantum size effects. A schematic of various techniques is shown in Fig. 1. However, the challenge comes in making the lithographic dimension the same as the confining potential dimension of the electron system.

The first technique that comes to mind to create the confining potential is brute force anisotropic dry etching. The basic principle is to use energetic ions to either erode or chemically react with the epitaxial material structure. By using a reactive gas species which form volatile compounds with the material, semiconductor structures as small as 20nm have been demonstrated. Thus, hard wall potentials can be formed by etching either partially or completely through the epitaxial structure. However, there are two serious drawbacks to this approach. First, a serious side effect of dry etching is damage to the semiconductor by the energetic ions. The extent of the damage is poorly understood, and has shown large process variability. Second, the resulting free surfaces have both Fermi level pinning and a large concentration of non-radiative recombination sites. Thus, both optical and electronic transport studies are problematic. Only in two special arrangements has this technique yielded conclusive results. Thus, other avenues of fabrication have been explored.

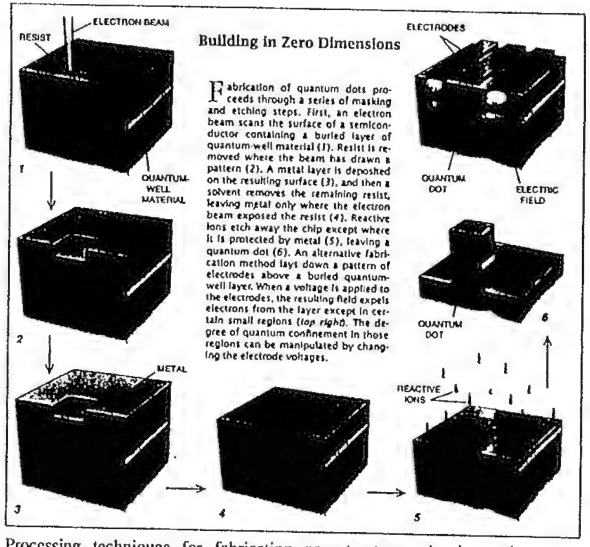


Figure 1. Processing techniques for fabricating nanostructures. A photoresist, such as PMMA (polymethylmethacrylate) covering a two-dimensional quantum well is exposed by a tiny beam of electrons. After developing, metal is evaporated to either serve as a mask for etching, implantation, or metal electrodes which squeeze electrons into the confinement regions.

A technique that bypasses the need for exposing critical surfaces is to define gates on the epitaxial structure which confine the underlying 2D system, often a two dimensional electron gas (2DEG). By applying a negative potential, the underlying 2DEG is depleted underneath the gates and confined to the region between the gates. This approach results in a smooth electrostatic confinement that has the advantage of tunability, unlike the cutting technique described previously. It also has the advantage that nanometer-size imperfections in the definition of these structures are not transferred to the confining potential, since the screening length is significantly larger than these imperfections. However, this also means that the size of the confined region is large (>100 nm) and relatively shallow (compared to heterojunction confinement), resulting in quantum states spaced by ~1 meV. Yet even with these limitations, this technique has proven to be quite successful in physics explorations.

An interesting problem which typifies the field is that these structures have non-local quantum interference due to relatively large coherence lengths. In these systems, contacts to the system are non-trivial - they are now by definition part of the entire (interacting and interfering) electron wave system. Thus, electrical leads become intractably invasive. What the physicist finds fascinating is for the device engineer a serious issue, as the ability to create an isolated gate circuit architecture is now problematic.

The holy grail of this field is the technological achievement of a confining heterojunction technology that allows low-dimensional confinement equal to that presently done in only the epitaxial dimension. Once a lateral heterojunction technology is achieved, subband spacings of tens, perhaps hundreds of meV will be realized, with resulting quantum

transport that is dominant and perhaps technologically useful. This degree of control, which has been achieved in the epitaxial dimension, may give rise to a host of new, promising electronic and optoelectronic devices. Although the focus of this chapter is on electron device technology, one should be cognizant of similar advances in low dimensional optoelectronic devices which may benefit from the same advances.

An exciting contender for advances in this regime is the achievement of low dimensional structures by in situ epitaxial growth. There exist a number of approaches for the realization of such structures, such as the overgrowth onto patterned substrates and ``cleaved edge overgrowth", where 2D MBE heteroepitaxial material is cleaved in situ, rotated, and subsequent growth occurs directly onto a 2D interface. These approaches are exciting although they have extreme technical challenges for useful device application.

To gain an appreciation of the technological possibilities of these technologies, with regard to room temperature operation and gain (i.e., large operating voltage), Fig. 2 presents a comparison of different quantum device technologies, plotted as a function of their experimentally observed operating temperature and voltage. At this time, only the technologies that utilize heterojunction barriers for charge separation and tunneling provide the requisite requirements. This figure is useful as a guide to lead us toward fabrication technologies that will enable potential quantum and mesoscopic devices. This chapter reviews some of the various fabrication and materials technologies for making nanoscale devices and structures.

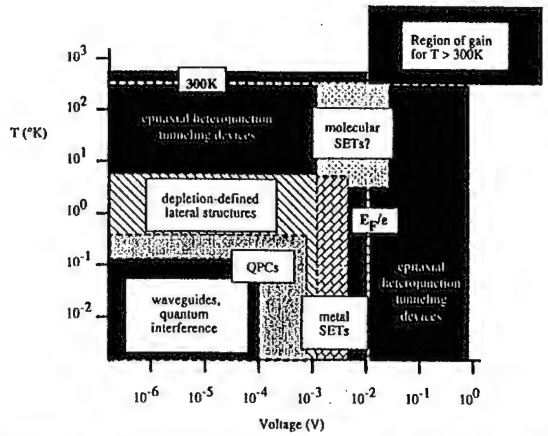


Figure 2. Comparison of different quantum device technologies, plotted as a function of their experimentally observed operating temperature and voltage. ``QPCs" denotes quantum point contact structures; ``SETs" to Single Electron Transistors. Room temperature and a voltage equal to the Fermi level voltage are noted.

NANOLITHOGRAPHY

There exists a wide and continually expanding variety of methods to perform lithography on the scale here of interest for nanoscale electronic devices, which is here defined as sub-0.1 micron feature size. Electron beam lithography, focused and masked ion beam lithography, x-ray lithography, and more recently scanning probe microscopies used in lithography mode are the most common techniques used today, though by no means the only techniques. Figure 3 compares the resolution limits of these techniques. Clearly conventional optical lithography is unsuitable for creating features in this regime, due to wavelength and depth of focus limitations. Although ``slight-of-hand" tricks such as step-edge shadowing and holography can be used to create lines and/or periodic structures that extend into (and in some cases considerably below) the sub-0.1 micron regime, they do not have the topographic or processing generality needed for contemporary device structures.

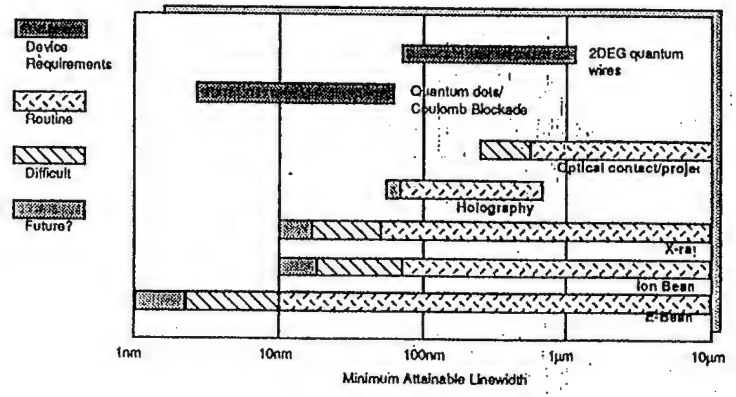


Figure 3. A comparison of the resolution limits of various lithography techniques.

By far the most common tool used for nanolithography today is electron beam (e-beam) lithography. In addition to superiority in demonstrated minimum feature size, e-beam lithography has advantages in ease of use, flexibility in pattern redesign, and alignment. Similar in concept are focused ion beam systems, which have the advantage of resistless implantation and/or etching. These are contrasted with X-ray and masked ion-beam lithography, and other novel approaches. Following we discuss the system, resist, and process issues for each of the techniques for fabricating ultrasmall devices.

Electron Beam Lithography

Electron Beam Lithography Systems. E-beam systems that are utilized for both nanolithography and mask generation are commercially available. However, these systems are generally not accessible to the average user due to the prohibitive multi-million dollar cost and (perhaps more importantly) intensive maintenance. Below we discuss conversion of a scanning electron microscope into a lithography tool that has nearly identical performance to that of a mask generation system, and specifically meets the needs of ultrasmall device researchers.

A schematic of a general e-beam system is shown in Fig. 4. The critical issues important to the nanofabricator are:

* Minimum feature size, of course. Surprisingly to the beginner, the minimum feature size is rarely a measure of the electron optics or spot size. For direct write onto thick substrates, the major contributing factors for minimum feature size are backscattering and resist limits. These will be discussed later.

* Source brightness. Resist speed is normally the limiting factor in system throughput, thus source brightness needs to be maximized. This is especially true for minimum feature (and therefore spot) size, when the electron beam current is often in the *pA* range. The value of increasing throughput is often neglected in non-mask making or dedicated full wafer direct-write applications. However, the range of possible device structures, process and exposure testing via step-and-repeat capability, and alignment capability are severely limited by reduced throughput. Tungsten wire "hairpin" filaments are standard for most SEMs, but when converting to lithography are often replaced by LB_6 crystal (luminosities ~ 5-10 times larger) or cold-cathode field emission (luminosities as great as 1000 times larger) filaments. The disadvantage of these sources is the additional cost of the filaments and the related high vacuum equipment.
* Stage movement and alignment. The fabrication of ultrasmall devices rarely has a yield that approaches unity, so the fabrication of "one-at-a-time" devices often produces failures one-at-a-time. Automated stage movement, simple to implement from numerous commercial sources, alleviates labor-intensive step-and-repeat and the throughput bottleneck. Micron-level blind movement produces large scale alignment, such as shown in Fig. 5, acceptable for bonding-pad or gross interconnect connection to critical feature structures. Automated sub-micron alignment is normally reserved for interferometric stages, although it is possible to implement on converted SEMs with care.
* Pattern generation. The implementation of a pattern generator is relatively straightforward, using commercially available graphic design programs, a PC controller with interfacing, and care in eliminating stray noise in the D/A stage. Some speed is sacrificed in implementing a bitmap pixel-writing method over superior vector writing, but can be minimized with software algorithms making the limiting speed essentially resist-limited. Some commercial systems will employ shaped-beam control, a significant speed improvement, but is beyond the capability of a conversion system.

Figures 5-8 show illustrative lithography results from a converted SEM. The system is a 40 KV, 3 nm spot size LaB_6 filament JEOL 6400 with magnetic beam blanking, an *x-y-z* computer-controlled stage, a 486 33 Mhz PC with a custom interface board (to enhance throughput) outputting to 16-bit D/As for mastered pixel-writing. Figure 5 and 6 show lithography results (after metal lift-off) illustrating step-and-repeat capability for a test pattern; note the intentional exposure variations of Figure 6 used for determining correct exposure parameters.

Figures 7(a)-(e) show an exposure series of the same test pattern in the high resolution region of the Fig. 6 fields, illustrating a number of exposure, proximity, resist, and metal transfer issues. First, one can see that the self-proximity (intraproximity) effect of each structure determines critical exposure; both geometry and size are important. For all patterns shown here, the same pixel size, spacing, and dose was used. (Obviously complex patterns demand tailoring of these parameters for correct pattern exposure). Second, proximity effects between patterns (interproximity) are evident (best illustrated by the horizontal line series and the arrowhead). Third, note that small enclosed geometries do not give acceptable resist profiles for lift-off, due to sloping resist profiles, leaving residual metal in place or nearby. The use of these test patterns allow one to fine-tune the exposure parameters, often critical to a few percent, for complex geometries. Figure 8 shows various mesoscopic structures with 2nm minimum features, after lithography and pattern transfer into a 2DEG illustrating proper dose control and tailoring to reduce proximity.

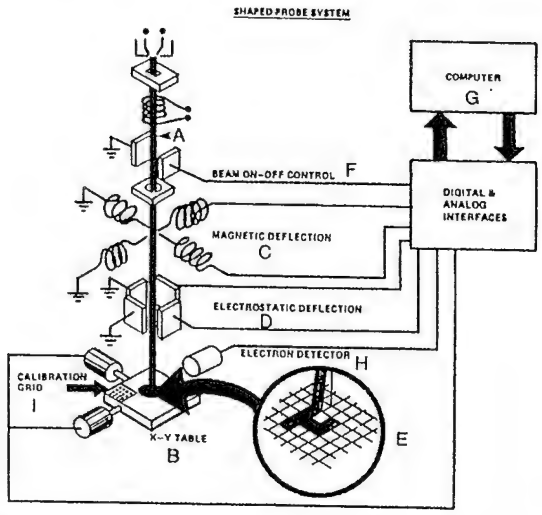


Figure 4. Schematic of a general e-beam system.

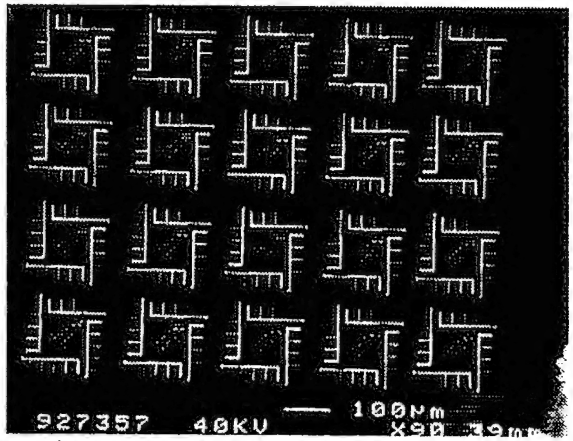


Figure 5. Demonstration of step-and-repeat (after metal lift-off).

Electron beam resists. Positive radiation resists are required for ultimate resolution for nanometer high-energy ebeams. The highest resolution high-energy radiation resist resist is polymethylmethacrylate (PMMA), in which the long polymer chains are sissioned into smaller fragments that have a higher differential solubility than the unsissioned chains. Figure 9 shows a comparison of contrast ratios that can be achieved in monolayer PMMA. Common PMMA resists used vary from 100K to 950K molecular weight (MW), often using methylisobutylketone (MIBK) as a solvent. These resists have sensitivities ranging from 10-

100 microCoulombs/cm^2, and gammas (contrast ratios) of >3 in MIBK/IPA (isopropyl alcohol) developers. Bilayers or multilayers are also used and have similar developing properties. In the case that multiple PMMA layers are used, the lower(est) MW weight PMMA is placed below (on the bottom) of the layered stack, since the solubility decreases with increasing molecular weight.

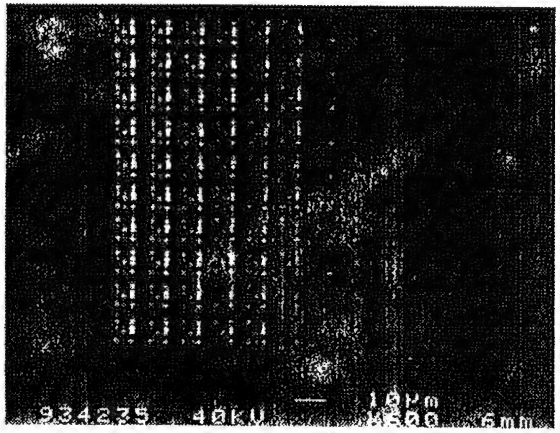
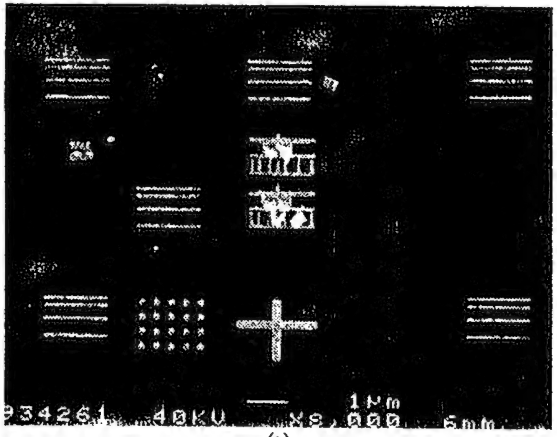
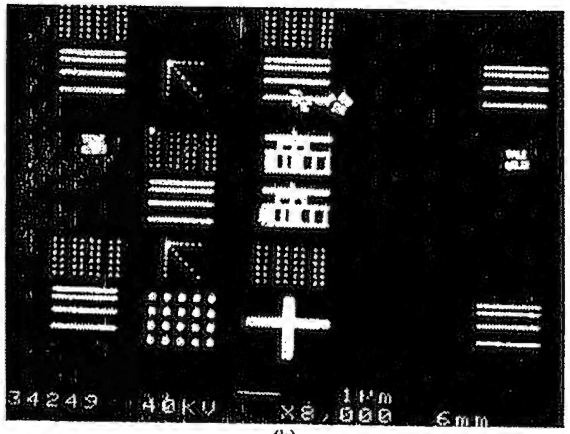


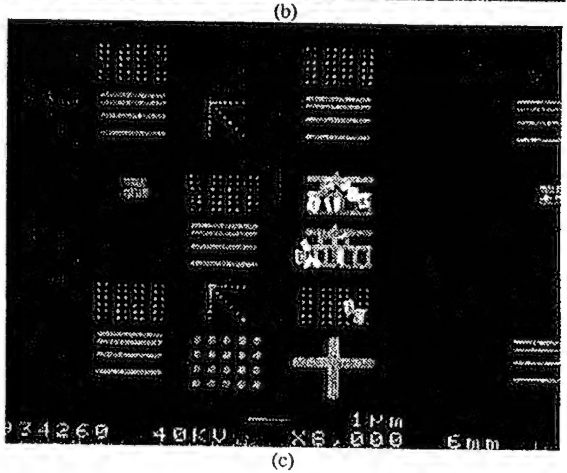
Figure 6. Exposure parameter series to determine correct parameters. The dose varied from 214 microCoulombs/cm^2 to 10,000 microCoulombs/cm^2, using a 2 nm pixel and 3.6 pA of current at 40 KV. Si substrate, after liftoff.



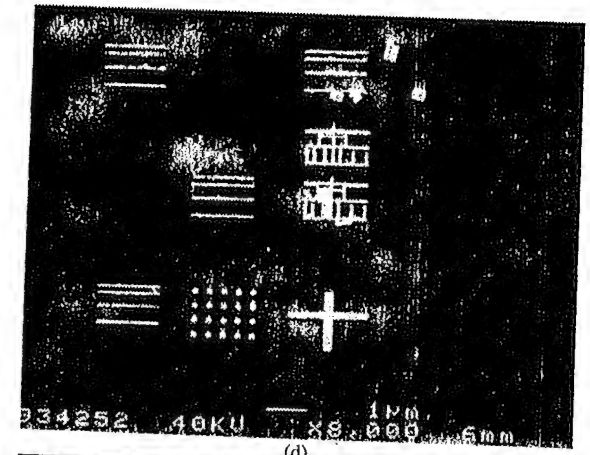
(a)



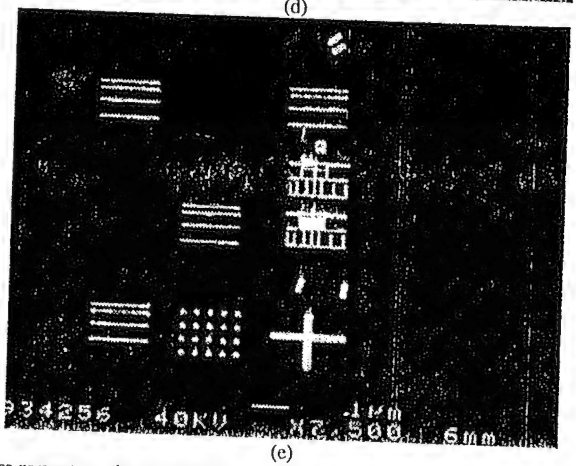
(b)



(c)



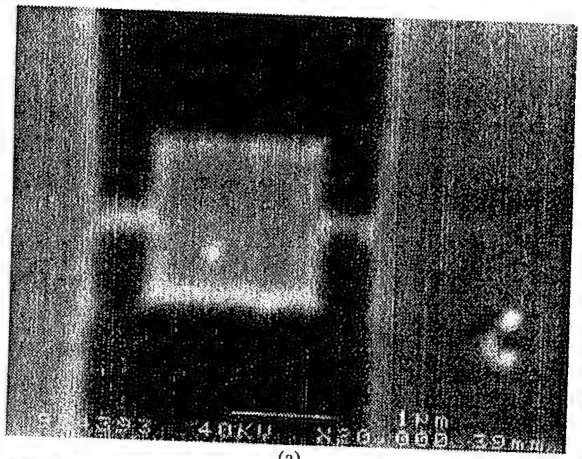
(d)



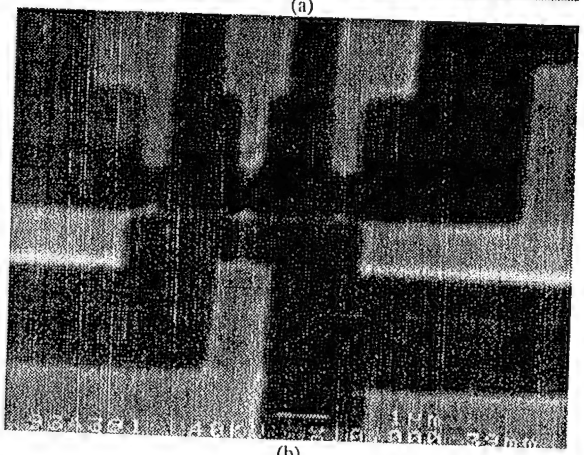
(e)

Figure 7. Dose parameter series, magnified views of Fig. 4. The doses used were (a) 1500, (b) 1575, (c) 3600, (d) 3800, and (e) 6045 microCoulombs/cm^2. All used a a 2nm pixel and 3.6 pA of current. The lifted-off metal was 5 nm of Ti with 35 nm of Au.

The ultimate theoretical resolution of PMMA is still a matter of some controversy; regardless, it has been observed that 10nm is the practical limit for linewidth definition. In contract, negative resists (which utilize low MW polystyrene derivatives that crosslink upon irradiation) have minimum demonstrated resolution of 30 nm. To advance beyond the 10 nm barrier in resist, there has been work in various inorganic resists that are either directly dissociated or ablated, such as strontium fluoride or magnesium chloride. These resists have demonstrated features down to 2 nm. However, the transfer of these resist patterns to useful device structures without degradation of resolution has yet to be demonstrated. As of today, PMMA is still the highest resolution useful resist for pattern transfer.



(a)



(b)

Figure 8. Various mesoscopic transport structures. The dark regions are etched GaAs.

Negative resists, although having poorer nanometer-scale resolution, are still extremely useful in some applications. An example if for >30 nm minimum resolution clear field masks or direct write patterns. As will be discussed later, the critical minimum dimension of many device structures is not limited by the minimum attainable lithographic dimension but by device issues that imply significantly larger dimensions. For these applications, maximum resolution versus other factors such as speed and ease of use are important.

Proximity. The curse of e-beam lithography is the proximity effect. The backscattering of secondary electrons from the substrate produces a diffuse lateral background of exposure that causes extensive resolution degradation, especially in the 10-

30 KV range. Figure 10 illustrates the effect of backscattering. Proximity exposure can originate within a pattern itself (intraproximity) or from adjacent patterns (interproximity). Both effects are illustrated by example in Fig. 7.

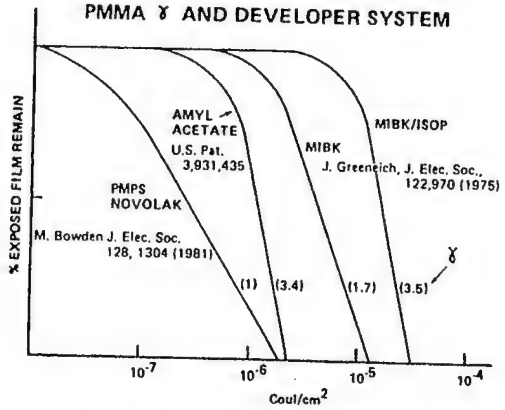


Figure 9. A comparison of contrast ratios that can be achieved in monolayer PMMA.

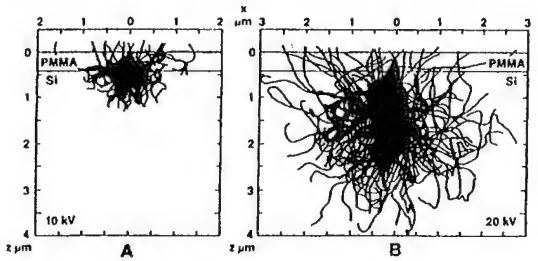


Figure 10. Proximity effect by backscattering of secondary electrons.

There are a number of techniques to reduce this effect:

* Eliminate the substrate. By etching the substrate to a thin membrane (such as depositing Si3N4 on a Si wafer, etching away the Si wafer from the back, then depositing PMMA and exposing), ultimate PMMA limits can be achieved. However, this is obviously of limited utility for device applications.
* Use accelerating voltages of 50-100KV or greater.
* Use multilayer resists. As mentioned above, one can create a bilayer PMMA with a bottom lower MW weight layer. This bilayer increases the substrate to defining layer distance, as the lower MW is essentially a spacer and sacrificial layer. Subsequent development also produces an undercut profile. This can improved upon by a trilayer (or multilayer), where the center layer is a high-Z metal (such as Au) which increases the intensity of electrons in the top defining resist (i.e., effectively increasing the sensitivity) due to the larger backscattering coefficient in comparison to the substrate. Alternatively, the middle layer can simply be a delimiter layer between the top and bottom resist layers, of a material easily compatible with anisotropic etching (such as Ge). This allows for high fidelity pattern transfer to a very thin top layer.
* Tailor the pattern dosage to correct for proximity. or any combination of the above.

It should be note that backscattering is of course a function of substrate composition, and patterns optimized for a given substrate will in general change significantly on a different substrate. Figure 11 shows that same pattern, again at constant pixel dose for a GaAs substrate (Fig. 7 is for Si substrates). The dose here was 3200 microCoulombs/cm^2, approximately 2× that for the appropriate Si substrate dose. One can notice that interproximity effects are different; for example, note the fidelity of the lettering as compared to the interproximity-determined horizontal linewidths.

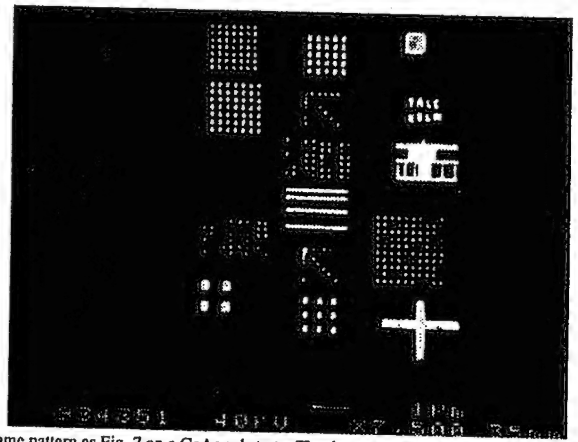


Figure 11. Same pattern as Fig. 7 on a GaAs substrate. The dose was 3200 microCoulombs/cm^2.

Diagnostics. The long-standing diagnostic tool of the nanofabricator has been lift-off of metal. Recently, atomic force microscopy (AFM) has developed as an extremely useful diagnostic of lithography in this regime. The advantage of this technique is that it non-destructively examines the resist profile prior to subsequent processing, such as lift-off, which often masks resist profile and/or exposure problems due to introduction of another processing step. Figure 12 shows an AFM of a resist pattern similar to Figure 8(b) but with a trapezoidal center region and an intentional "notch". Figure 12(a) shows the pattern for a properly defined pattern, and Fig. 12(b) a top view. A similar and nominally identical pattern is shown in Fig. 12(c), showing irregularities in the pattern due to "shifted" pixels (note the dark spots on the left). Such irregularities would give ambiguous pattern-transfer results. (Here the origin of the irregularities were subsequently traced to a mechanical instability in the filament.) Likewise, Fig. 12(d) shows a correct exposure in the device region but an overexposure in the top and bottom contact regions, giving sloping resist profiles. Subsequent lift-off would also be ambiguous, with alternative explanations including resist adhesion and thickness.

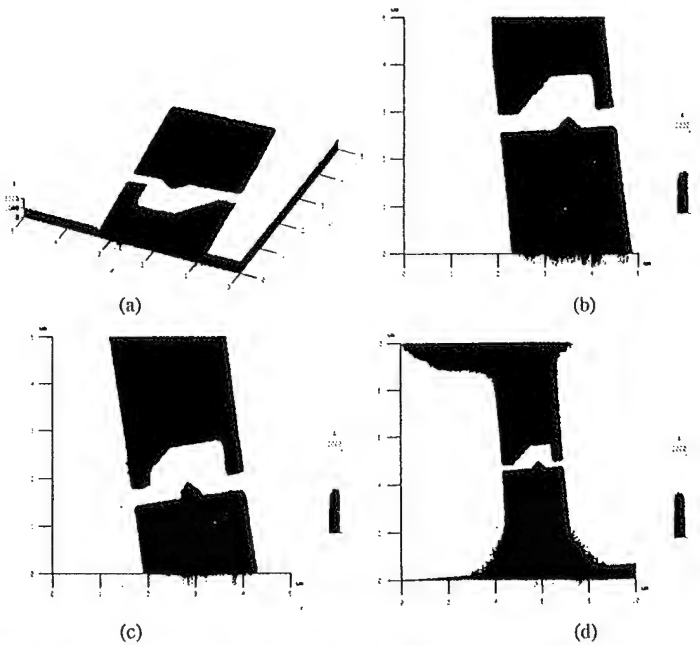


Figure 12. An AFM of a resist pattern similar to Fig. 8(b) but with a trapezoidal center region and an intentional "notch". (a) The pattern for a properly defined pattern, and (b) a top view. (c) Irregularities in the pattern due to "shifted" pixels due to mechanical instability in the filament. (d) Overexposure in the top and bottom contact regions. Note the sloping resist profiles.

Ion Beam and X-ray Lithography

Ion beam exposure comes in two flavors; 1) as a projection printer, and 2) as a focused beam similar to ebeam. The main advantage to using ion beams in either case is the lack of backscattering and hence proximity effects. Additionally, ions are absorbed 10-100× more efficiently than electrons, increasing the effective sensitivity of the resist by that amount. For nanometer applications, this may even be a problem due to shot noise limitations.

As a projection printer, stencil masks are proximity gapped except in submicron work where the mask is used in contact. In this case, complementary masks are used to avoid the doughnut-hole problem. Fig. 13 shows various stencil mask techniques used. Although the technique shows very promising results, it suffers from problems of compact sources and mask fabrication/stability. Because of these, eventual large field exposure is problematic due to penumbra effects and wafer warpage. X-ray lithography suffers from similar problems, in addition to having less sensitive resists.

Focused ion beams have an additional advantage for resistless processing, mask repair, and direct etching in addition to resist exposure. Ion sources have improved considerably, with Ga, In, H, Si, P, and B being some of the available sources. However, in addition to significant capital expense, beam current is a liming problem. The difficulty is not in source brightness, but in total flux from the source. This is due to classical space charge spreading, which increases with m/q. Comparisons of throughput for a focused ion beam system versus an ebeam system based on source current, even taking increased resist sensitivity into account, shows the ion beam system ~300 times slower.

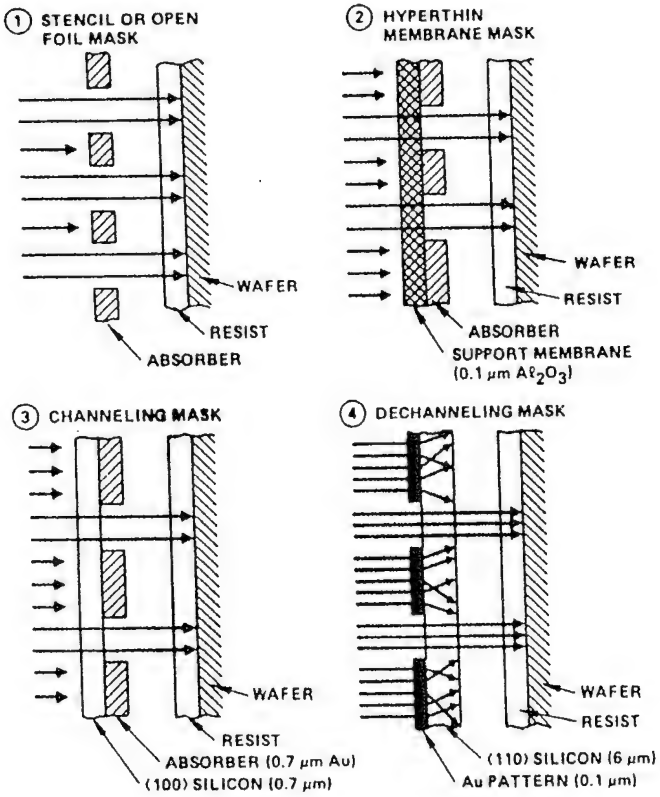


Figure 13. Various ion beam stencil mask techniques.

STM Lithography

The availability of an atomically-sharp point source at the substrate, at low voltage that clearly has no backscatter problem, makes STM lithography appear to be a clearly superior method. In general, this is not true. The main reasons are:

* To make a useful pattern transfer (in most cases), the resist thickness elevates the tip significantly above the substrate, changing the configuration from a single-atom-to-substrate to an extended source with fringing field.
* Ultimate resist limitations. STM lithography on polymeric resists has not demonstrated results superior to ebeam, and in general typically worse. The

additional disadvantages of field-of-view, stage motion, substrate topography problems, and speed limits STM as a general lithography tool. An advantage lies in the substantially lower system cost, making nanolithography available to most researchers with the mentioned limitations.

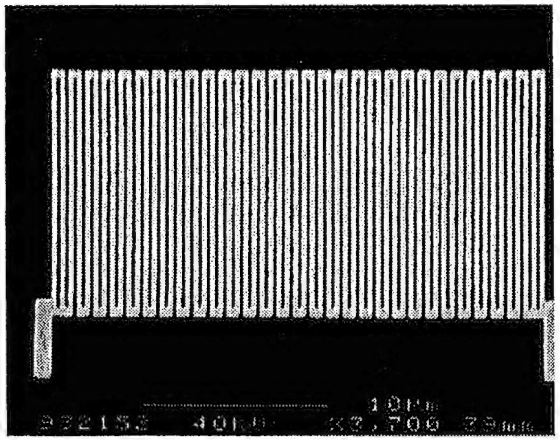


Figure 14. An illustrative example of atomic-level lithography.

However, without polymeric resists STM lithography is unmatched in resolution, which is atomic. This is in systems where the "resist" is a single atomic layer, for example H on Si or various adatoms on metal surfaces. Albeit not general purpose as yet, the technique demonstrates that atomically sharp definition can be done. Figure 14 shows an illustrative example of atomic-level lithography. This is a rapidly advancing field, where numerous avenues of exploration in resist (self-assembled monolayers, LB films, various adsorbate on atomically-sharp surfaces in UHV, etc), system (various combinations of scanning microscopy techniques, such as near field optical for one), and pattern transfer await exploration. The development of a general purpose tool for wafer-scale fabrication is problematic, though not intractable.

PATTERN TRANSFER

Pattern transfer masks

For nanolithography, most pattern transfer begins with metal evaporation and lift-off. Figure 15 illustrates this method. The metal or any other can be deposited by thermal evaporation. The resist acts as a stencil mask,which is washed away along with the unwanted material in an appropriate solvent. This technique has been used to deposit 10nm metal lines. For nanometer applications, source size and sample distance can often be limiting.

The thickness of the deposited metal is, at best, the thickness of the resist. The deposition of thick metal demands multilayer resists. The pitch of defined structures is limited by bottom layer undercut. Figure 16 illustrates the spatial uniformity and pitch that can be achieved with optimization. Here the metal thickness is 1/4 micron, with equilavent pitch.

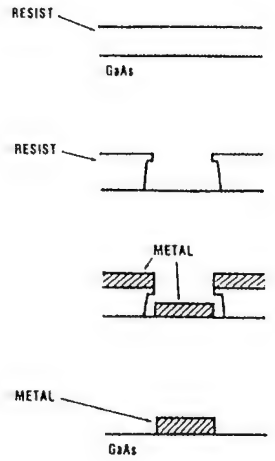


Figure 15. The lift-off technique.

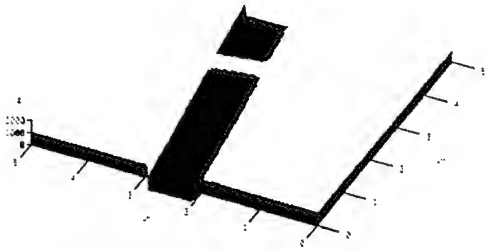


Figure 16. An example of a small pitch thick metal pattern using multilaterresist.

Shadow evaporation techniques are an interesting twist in nanometer fabrication. In this technique, shown in Fig. 17, the source metal position is changed with respect to the substrate at various points in the deposition. Coulomb blockade devices, which require small interelectrode capacitance and an intervening insulator layer, have very successfully employed this technique. Figure 18 shows an example of using shadow evaporation for the formation of ~10 nm gaps between electrodes; (a) shows an AFM of the top resist profile. What cannot be seen from the AFM image is that the resist is actually a bilayer, where the bottom spacing layer under the bridge has been selectively developed away. This forms a free-standing bridge which can be used for shadowing evaporated metal. (b) and (c) show the resultant lifted-off metal pattern, where metal from two source directions have been deposited (~ 10 degrees apart). The gray-scale in the image indicates the metal thickness. This is a surprisingly reliable and high yield technique.

Resists can also be used directly for subsequent pattern transfer, such as an etch mask or for topographic definition of a gate by evaporating metal directly onto the resist and leaving the resist in place.

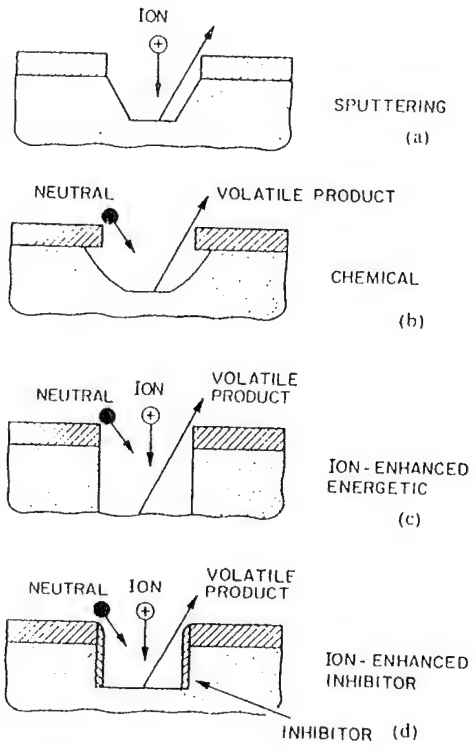
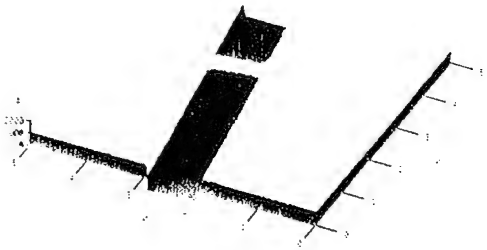
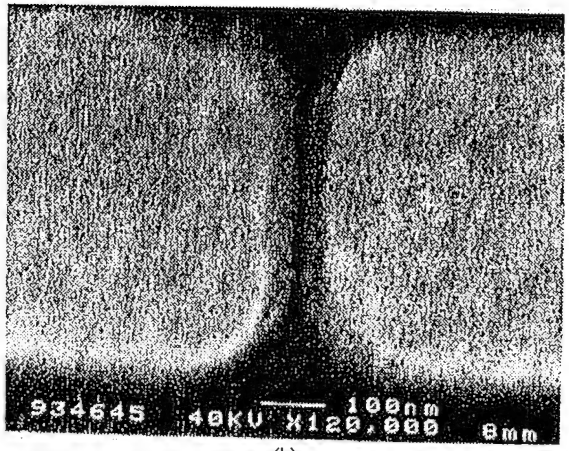


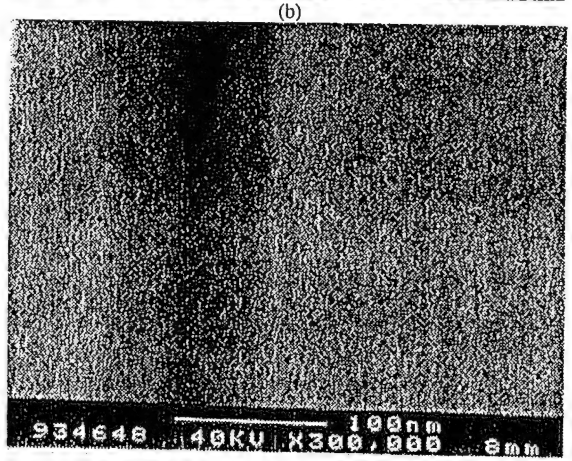
Figure 17. Shadow evaporation.



(a)



(b)



(c)

Figure 18. An example of using shadow evaporation for the formation of ~10 nm gaps between electrodes. (a) An AFM of the top resist profile. The bottom layer undercut cannot be seen. (b), (c) Resultant lifted-off metal pattern. Metal from two source directions have been deposited (~ 10 degrees apart). The gray-scale in the image indicates the metal thickness.

Etching

There are four basic mechanisms of plasma etching: a) sputtering, where the interaction is purely mechanical, with impingent high energy ions ejecting substrate surface atoms; b) chemical, where active species form a volatile product when reacting with the surface; c) ion-enhanced energetic, where a neutral species reacts with the surface in the presence of impingent high energy ions; and d) inhibitor-driven ion-assisted etching, where

the neutral species reacts spontaneously with the surface, and the role of the impingent ions is to remove the inhibiting layer. Figure 19 shows a schematic of these processes.

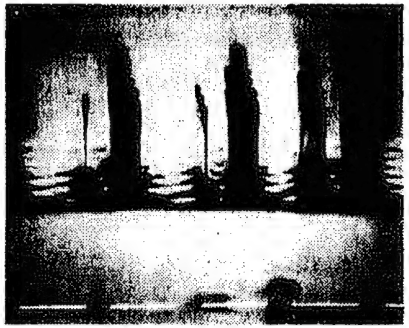


Figure 19. Basic mechanisms of plasma etching.

It should be emphasized that any combination of these processes can be created in a given system. The vast array of etching techniques, systems, and chemistries is far too extensive to survey here, but nonetheless all the techniques involve tailoring the four mechanisms above to the specific application. In nanometer applications, the important issues are often high anisotropy and low damage.

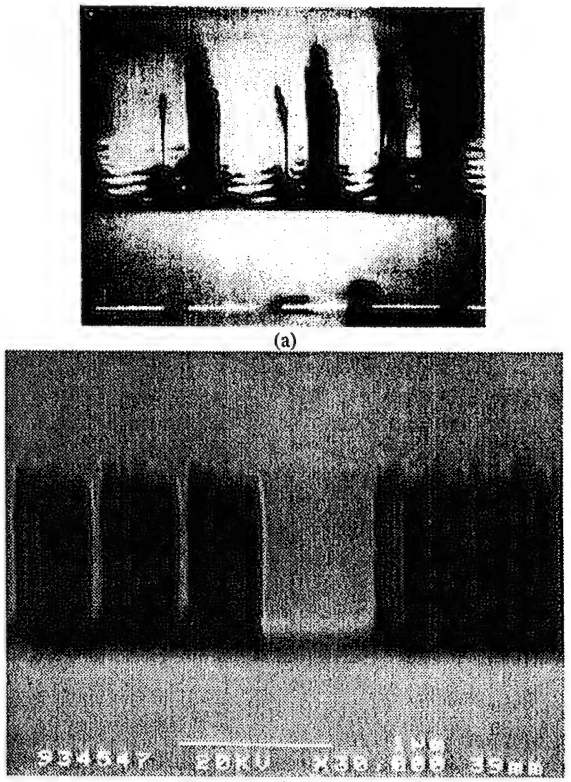
Table 1 shows some of the source gases and materials that can be etched/are resistant with Cl-based chemistry for some common material systems of interest. (F is also extensively used, especially for Si etching). Table x shows the effect of introducing various additives to the source etch gas, and the primary mechanism for their effect.

Very high anisotropy in etching can be achieved, retaining the nanometer lithographic lithographic dimension. Figure 20(a) shows a result of a GaAs pattern transfer using Cl-based RIE (reactive ion etching). As an example of choosing the proper technique (therefore mechanism) for a given application, compare this result to a CAIBE (chemically-assisted ion beam etching) result (Fig. 20(b))of a similar pattern also in GaAs. In this application integrity of the top mask is important; here the additional sputtering component has slightly eroded the top etch mask.

DEVICE ISSUES

Nanometer scale fabrication techniques, combined with epitaxial resonant tunneling structures, now routinely allow the study of quasi-0D confined electron systems. In addition to energy level separations that are tunable by the confining potentials, these systems can also exhibit Coulomb blockade. An extensive literature exists for investigation of quantum dots and Coulomb blockade in dots and metal structures, which will not be repeated here.

However, there are 4 issues that need addressing before significant advances can be made in the field of ultrasmall quantum/charge effects and devices. The first is the role of depletion layers in quantum semiconductor structures. In examples, such as that shown in Fig. 21 of a gate-controlled quantum dot, one immediately notes that the lithographic dimensions of the structure far exceed the minimum attainable resolution discussed in this chapter. This is because the imposed potential on the active region (in this case an underlying 2DEG) is dominated by the depletion length of the device. For many of the structures, this limits the lithographic dimensions to ~1/4 micron, with a soft parabolic confining region. Energetic splittings of ~1 meV or less are typical. For even the best confinement configurations, ~ 25 meV are the maximum attainable (with the exception of cleaved edge overgrowth). The role of depletion layers in present confinement schemes is dominant, and and alternative needs to be found before useful device structures are realistic. An entire exciting area of physics enabled by lateral heteroepitaxial confinement is waiting to be explored. The in situ growth of confined structures is an active area.



(a)

(b)

Figure 20. Anisotropic etching. (a) RIE of GaAs. (b) CAIBE of GaAs.

Second, fluctuations in both fabrication and materials become increasingly important with decreasing size scale. Devices depending on dimensional tolerance for operation, such as energy quantization, need ~10 kT energetic separation of states, which implies monolayer precision for threshold control. This clearly rules out existing fabrication schemes which have ~10 nm resolution limit. Even epitaxial structures have difficulty in achieving monolayer control, seen in both small scale and large scale fluctuations; thus, various overgrowth schemes will have similar difficulty unless self-limiting techniques are found. Coulomb blockade structures have their own fluctuations problems, such as background charge. Size fluctuations are also important for Coulomb blockade structures, both for

projected higher temperature operation (which implies ~1-2 nm dimensions) and the controlling gate potential which depends on the magnitude of a local capacitance versus a distribited standard. Precision in fabrication at the atomic scale is needed, which may involve techniques such as self-assembly of monolayers.

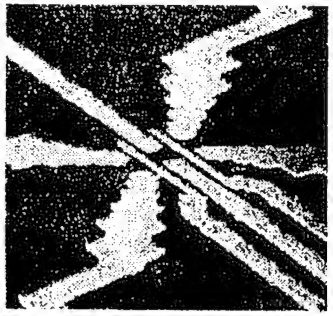


Figure 21. Metal gate electrode pattern of a quantum dot (on top of an underlying 2DEG). The spacings between electrodes confining the dot is 200nm. (Courtesy of R. Behringer, AT&T Bell Laboratories).

Third, the role of impurities are increasingly important. As the size of the structure decreases, the transport of a quantum or mesoscopic device can be dominated by a single unintentional impurity, or even the statistical distribution of intentional dopants. These effects can mask or complicate the measurement of quantum and/or Coulomb states in mesoscopic devices. Almost all of the fabrication processes outline above exacerbate this problem.

Finally, for useful electronic devices there is the issue of gain. Three terminal nanometer quantum devices with fanout and input/output isolation have yet to be experimentally achieved even in the laboratory.

CONCLUSION

The development of nanolithography tools has now led to exciting new discoveries in the one- and zero-dimensional regime. In addition to the enticing science of the low-dimensional regime, projected technological applications include designer materials, efficient lasers and optical amplifiers, and ultrasmall electron devices. Before these can be realized however there are a number of obstacles that have to be overcome.

The most challenging aspect is the unprecedented control over dimension and purity that must be achieved. There is a wide variety of fabrication techniques upon which to draw, but the existing techniques that are widely applicable reach resolution limits in the ~10 nm regime. Specialized demonstrations exist beyond this, and while are highly successful in exploring fundamental physics, are not generally applicable tools. The ``top-down'' approach to fabrication by carving, dicing, or squeezing semiconductors may not be able to achieve the requisite control for ultrasmall devices unless revolutionary advances in materials and nanofabrication are made. Precision in fabrication at the atomic level, and coincident purity, appears to be a requirement for future device structures. This constraint may necessitate fabrication techniques and materials quite different from present ones.

The present prototype device designs are large (though the active region of the device is quantum-sized, the defining electrodes and contact pads take up enormous space), and operate only at very low temperature. In addition they are made with fabrication technology that is not extendable to large scale integration. New lithographic tools which allow three dimensional atomic scale control, such as with structured epitaxial growth or those utilizing self-organization, are needed. The ``bottom-up'' approach with semiconductors demands ingenuity in surface chemistry. The solutions may force us to consider inventive material systems and synthesis techniques, or perhaps blends of conventional semiconductor technology with alternative approaches.

Finally, although there will be a realm of fascinating mesoscopic and nanoscale physics to explore, the end application to devices requires gain/fanout and room temperature operation.

Acknowledgements: We would like to thank M. Amman for the AFMs, and the support of ONR N00014-91-J-1561 and NSF DMR-9112497.

TRADITIONAL MODELLING OF SEMICONDUCTOR DEVICES

Christopher M. Snowden

Microwave and Terahertz Technology Group,
Department of Electronic and Electrical Engineering,
University of Leeds, Leeds. LS2 9JT. UK

INTRODUCTION

The study of semiconductor devices spans less than 50 years and the evolution of modelling techniques occupies an even shorter interval. Traditionally, the electrical properties of semiconductor devices have been modelled using equivalent circuit models, where the electrical behaviour of the device at the connecting terminals is represented by a circuit consisting of linear and non-linear circuit elements. However, this type of model can only give limited insight into the physical behaviour of the device and is often restricted in its application to well characterized devices. In contrast, physical models which describe the device in terms of the carrier transport properties and the material and geometrical attributes, allow both a physical and electrical description of the device. Physical models are by their nature more complex than equivalent circuit models and normally require numerical methods to obtain solutions to the set of transport equations. Physical models have generally been regarded in the past as requiring very powerful computers and are often considered to be the domain of the academic. Advances in computer technology over the past 15 years and powerful numerical methods have allowed physical models to address complex device structures using widely available computer resources. Simulation software capable of representing devices in two- and three-dimensions is now available commercially for use on workstations and is utilized increasingly in the semiconductor industry. The demand for supercomputer facilities is diminishing for these classical physical models.

The majority of semiconductor devices can be modelled using classical methods, where the influence of quantum phenomena are assumed to be negligible. Generally speaking, this includes all devices where the critical dimensions are significantly greater than a de Broglie wavelength. Examples of contemporary semiconductor structures which can be accurately represented using classical methods are: *pn* junction diodes, bipolar junction transistors, heterojunction bipolar transistors with layer thicknesses of greater than 0.1 microns, triacs, thyristors, MOSFETs and MESFETs with gate lengths of greater than 50nm. In the case of structures which support strong quantization it is necessary to solve at least Schrödinger and Poisson's equations self-consistently. At the present time the best example of a device requiring this type of treatment is the high electron mobility transistor (HEMT), where the charge-control model requires at least a one-dimensional solution of the Schrödinger-Poisson equations to describe the variation of charge distribution with electrostatic potential.

This chapter will provide an overview of the current state-of-the-art in traditional device modelling, with particular reference to two- and three-dimensional simulations. In order to appreciate the significance of recent advances it is necessary to have a sound appreciation of the basis of these models. Hence, an outline of the principles behind traditional semiconductor device models will be given in terms of a set of semiconductor equations which may be related to the Boltzmann Transport Equation. This will be followed by a discussion of other factors which influence the accuracy and realism of physical models, such as trapping phenomena in compound semiconductors and thermal effects. The chapter will conclude with an overview of several contemporary models for current semiconductor devices. The relatively advanced state of development of traditional physical models means that a high degree of quantitative agreement between simulated and measured data is now possible and consequently physical models are increasingly finding a role in computer aided design applications.

THE CLASSICAL PICTURE

The classical approach to characterizing carrier transport phenomena is to consider the carriers as charged particles with external forces acting on them. This can be treated from a purely phenomenological viewpoint or more usefully by considering a more fundamental physical description. The most common derivation of transport equations is based on making a series of approximations to the Boltzmann Transport Equation (BTE), which describes the evolution of a distribution function which in turn describes the properties of the charged particles.

The Boltzmann Transport Equation

Charge particles can be characterised in terms of their position in space **r** and momentum **k** at a time t. The density of charge particles $n(\mathbf{k},\mathbf{r},t)$ can in turn be described in terms of a distribution function $f(\mathbf{k},\mathbf{r},t)$, where

$$n(\mathbf{r},t) = \int f(\mathbf{k},\mathbf{r},t)d\mathbf{k} \quad , \tag{1}$$

The implicit form of the Boltzmann Transport Equation requires that the temporal derivative of the distribution function $f(\mathbf{k},\mathbf{r},t)$ vanishes along a particle trajectory **r**,**k** such that

$$\frac{d}{dt} f(\mathbf{k},\mathbf{r},t) = 0 \quad . \tag{2}$$

This is more easily interpreted by further expansion

$$\frac{\partial f}{\partial t} + \frac{\partial f}{\partial \mathbf{k}} \bullet \frac{\partial \mathbf{k}}{\partial t} + \frac{\partial f}{\partial \mathbf{r}} \bullet \frac{\partial \mathbf{r}}{\partial t} = 0 \quad . \tag{3}$$

Each of the three terms which constitute this expanded form can be interpreted in terms of the classical behaviour of a group of particles. The third term will be recognised as including the group velocity of the particles $\mathbf{v} = d\mathbf{r}/dt$. The second term describes the influence of the total force **F** acting on the particles, and is due to components associated with the internal crystal lattice $\mathbf{F}_I$ and external forces attributable to electromagnetic fields $\mathbf{F}_E$. This may be expressed classically as

$$\mathbf{F} = \hbar\frac{d\mathbf{k}}{dt} = \mathbf{F}_I + \mathbf{F}_E \quad , \tag{4}$$

where $\hbar$ is Planck's constant ($\hbar = h/2\pi$). The effect of the internal forces acting on the distribution function can be expressed statistically by describing the internal collision mechanism $(\partial f/\partial t)_{coll}$ as a function of a scattering probability. Hence, the second term in the Boltzmann Transport Equation can be expressed as

$$\begin{aligned}\frac{\partial f}{\partial \mathbf{k}}\bullet\frac{\partial \mathbf{k}}{\partial t} &= \frac{\mathbf{F}_E}{\hbar}\bullet\frac{\partial f}{\partial \mathbf{k}} + \frac{\mathbf{F}_I}{\hbar}\bullet\frac{\partial f}{\partial \mathbf{k}} = \frac{\mathbf{F}_E}{\hbar}\bullet\frac{\partial f}{\partial \mathbf{k}} + \left(\frac{\partial f}{\partial t}\right)_{coll} \\ &= \frac{\mathbf{F}_E}{\hbar}\bullet\frac{\partial f}{\partial \mathbf{k}} + \int d\mathbf{k}'\left[f(\mathbf{k}')P(\mathbf{k}',\mathbf{k}) - P(\mathbf{k},\mathbf{k}')\right]\end{aligned} \quad , \tag{5}$$

where $P(\mathbf{k},\mathbf{k}')$ is the probability of a particle being scattered and changing its wave vector from k to k'. This latter statistical interpretation of the scattering processes forms the basis for Monte Carlo particle simulation method for modelling carrier transport, which will be discussed elsewhere. Collecting the expansions in equations (4) and (5) yields are more easily understood form of the BTE

$$\frac{\partial f}{\partial t} + \frac{\mathbf{F}}{\hbar}\bullet\frac{\partial f}{\partial \mathbf{k}} + \mathbf{v}\bullet\frac{\partial f}{\partial t} = \left(\frac{\partial f}{\partial t}\right)_{coll} \quad . \tag{6}$$

Although there have been solution methods developed for the Boltzmann Transport Equation, there are no closed-form solutions available. It is necessary at this stage to make a number of approximations to simplify the model. Firstly, the distribution function is assumed to be symmetrical in momentum space. Secondly, the collisions are usually assumed to be instantaneous (or at least of very short duration compared with the intervals between collisions). Scattering events are assumed to be elastic and the probability of scattering is assumed to be independent of external forces. The interaction between particles is assumed to be negligible. The model is based on band theory and the effective mass principle. Degeneracy effects are neglected in this simple model. The external force is assumed to be attributable solely to the electric field E, that is $\mathbf{F}_E = -q\mathbf{E}$, neglecting the Lorentz force due to the magnetic field. Finally, the external forces are assumed to be constant over a distance comparable with the dimensions of the wave packet describing the motion of the particle. These approximations result in the following transport equation:

$$\frac{\partial f}{\partial t} - \frac{q\mathbf{E}}{\hbar}\bullet\frac{\partial f}{\partial \mathbf{k}} + \mathbf{v}\bullet\frac{\partial f}{\partial t} = \left(\frac{\partial f}{\partial t}\right)_{coll} \quad , \tag{7}$$

where q is the magnitude of the charge on the electron. The collision term is usually considered in terms of a relaxation time approximation with contributions from the carrier density n, carrier velocity v and carrier energy w,

$$\left(\frac{\partial f}{\partial t}\right)_{coll} = \left(\frac{\partial n}{\partial t}\right)_{coll} + \left(\frac{\partial \mathbf{v}}{\partial t}\right)_{coll} + \left(\frac{\partial w}{\partial t}\right)_{coll} = -\frac{f - f_0}{\tau} \quad . \tag{8}$$

Here, f_0 is the spherically symmetric solution and τ is the relaxation time. The collision terms are represented using generation-recombination rates, momentum and energy relaxation rates, as

$$\left(\frac{\partial n}{\partial t}\right)_{coll} = -G \quad , \quad \left(\frac{\partial \mathbf{v}}{\partial t}\right)_{coll} = -\frac{\mathbf{v}}{\tau_m} \quad , \quad \left(\frac{\partial w}{\partial t}\right)_{coll} = -\frac{w - w_0}{\tau_w} \quad , \tag{9}$$

where G is the generation-recombination rate, τ_m and τ_w are the momentum and energy relaxation times.

The Semiconductor Equations

The above approximation for the Boltzmann Transport Equation can be used to obtain sets of transport equations which describe the behaviour of carriers in terms of carrier density, velocity, mobility, energy, effective mass, diffusion and relaxation parameters. The most popular route for arriving at sets of transport equations is based on taking moments of the Boltzmann Transport Equation, which yields the *conservation* or *balance* equations. This process is detailed in several texts (for example: Snowden, 1988; Tomizawa, 1993). The popular derivation of these transport equations assumes that the carrier distribution is of a drifted Maxwellian form, with a single spatially homogeneous conduction band and spherical constant energy surfaces. If the first three moments of the Boltzmann Transport Equation are taken a transport model is obtained which yields the full form of the carrier (current), momentum and energy conservation equations. In the following discussion we will restrict the treatment to electrons. The hole transport equations can be obtained in a similar manner. The current continuity equation is obtained by integrating the BTE over momentum space.

$$\frac{\partial n}{\partial t} + \nabla\bullet(n\mathbf{v}) = -G \quad \textit{for electrons}, \tag{10}$$

where n is the electron density and G is the generation-recombination rate (positive for genberation). The momentum conservation equation for electrons is obtained by multiplying (7) by k and integrating over momentum space. Expressing the result in terms of the average carrier velocity v and average carrier energy w yields

$$\frac{\partial \mathbf{v}}{\partial t} = -\frac{q\mathbf{E}}{m^*(w)} - \frac{2}{3nm^*(w)}\nabla(n\mathbf{v}) - \frac{\mathbf{v}}{m^*}\nabla(m^*\mathbf{v}) + \frac{1}{3n}\nabla[m^*(w)v^2] - \frac{\mathbf{v}}{\tau_m(w)} \quad . \tag{11}$$

In the case of homogeneous structures, the effective mass m^* in the third term in (11) can be taken outside of the derivative. However, this form of momentum balance equation should be used in heterostructure simulations where the material parameters are spatially dependent. The energy conservation equation is obtained by multiplying (7) by the electron energy w and integrating over momentum space. Again, expressing the result in terms of average electron energy and average electron velocity, the energy balance equation is obtained as

$$\frac{\partial w}{\partial t} = -q\mathbf{v}\bullet\mathbf{E} - \mathbf{v}\bullet\nabla w - \frac{2}{3n}\nabla\bullet\left[\left(n\mathbf{v} - \frac{\kappa}{k}\nabla\left(w - \frac{m^*(w)}{2}v^2\right)\right)\right] - \frac{w - w_0}{\tau_w(w)} \quad , \tag{12}$$

where k is Boltzmann's constant and κ is the thermal conductivity of the semiconductor material. Here the terms multiplying the κ/k coefficient represent the heat flux by assuming that the heat flow is proportional to the temperature gradient. The electron energy is given by

$$w = \frac{m^*(w)}{2}v^2 + \frac{3}{2}\kappa T_e \quad , \tag{13}$$

where T_e is the electron temperature. The equilibrium electron energy is given by

$$w_0 = \frac{3}{2} kT_0 \ , \tag{14}$$

where T_0 is the lattice temperature. The parameters $m^*(w)$, $\tau_w(w)$, $\tau_m(w)$, are determined from their relationship with the steady-state electric field E_{ss} obtained from Monte Carlo simulations (Carnez *et al.*, 1980; Snowden and Loret, 1987). The mobility, average electron energy, energy relaxation time and upper valley occupancy for GaAs obtained from Monte Carlo simulations are shown in Figure 1.

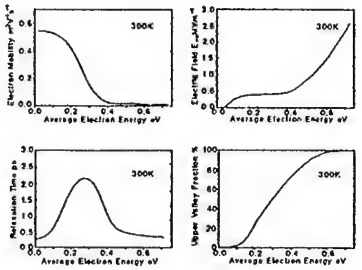


Figure 1. Steady-state transport parameters for GaAs. (a) electron mobility (b) electron energy (c) relaxation time (d) upper valley occupancy

These steady-state relationships vary significantly for different semiconductor materials and it is important to have a complete set of data to accurately model heterostructure devices (early heterojunction simulations often ascribed similar transport parameters to GaAs and AlGaAs). The energy-electric field relationships for GaAs. InGaAs and AlGaAs are shown in Fig. 2.

It should be noted that strictly speaking the assumption of a drifted Maxwellian distribution function leads to a set of transport equations which overestimate the thermal diffusion by a factor of 2 (the coefficient of the $\tilde{N}nw$ term in the momentum conservation equation) and yield zero carrier heat conductivity in the energy conservation equation. Carrier heat conductivity is introduced, as above, by applying the Weidermann-Franz law (Bosch and Thim 1974). A more recent treatment assumes spatially inhomogeneous sub-bands and ellipsoidal constant energy surfaces in association with a perturbed BTE solution (Marshak and van Vliet 1984) in place of the drifted Maxwellian approach (McAndrew et al 1987). This derivation overcomes the inaccuracies of the transport equations derived from a Maxwellian basis. The transport equations described above are for a single species of carrier (electrons) in one valley. Thus a complete model requires similar sets of equations for electrons and holes in other valleys. In practice, this is computationally prohibitive and a single-electron gas model with average is usually adopted (Blotekjaer 1970). The single-electron-gas approach accounts for inter-valley transfer effects and describes the transport parameters in terms of average electron energy.

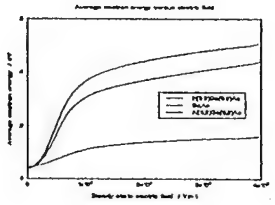


Figure 2. Steady-state energy-electric field characteristics for GaAs, GaInAs and AlGaAs obtained from Monte Carlo simulations.

The Poisson equation, derived from Maxwell's third equation, is generally used to relate the electric field to the charge

$$\nabla \bullet (\varepsilon_0 \varepsilon_r \mathrm{E}) = -q(n - p - N_D^+ + N_A^- - N_T^i) \ , \tag{15}$$

where ε_0 is the permittivity of free space and ε_r is the local permittivity of the material, N_D^+ and N_A^- are the local ionized donor and acceptor doping densities and N_T^i is the net ionized local trap density. It should be noted that this generalized form of Poisson's equation is required for heterostructure modelling, where the material layer structure is inhomogeneous in nature. The electric field E is related to the electrostatic potential ψ by

$$\mathrm{E} = -\nabla\psi \ . \tag{16}$$

Many simple models consider homogeneous domains, with constant parameter values, and in these circumstances the Poisson equation reduces to

$$\nabla \bullet \mathrm{E} = -\nabla^2\psi = -\frac{q}{\varepsilon_0 \varepsilon_r}(n - p - N_D^+ + N_A^- - N_T^i) \ . \tag{17}$$

In unipolar models it is only necessary to solve a sub-set of the transport equations, considering only the electron or hole continuity and current density equations, omitting generation-recombination effects. This approach was commonly chosen for many microwave devices such as Schottky diodes, MESFETs and HEMTs (for example, Reiser, 1973; Snowden *et al.*, 1983), although more recently features such as p^- buffer layers, breakdown and optical simulation (Tian *et al.*, 1992) of these unipolar devices have required full bipolar treatments. Bipolar devices, such as pn junction diodes, BJT's and HBT's intrinsically require a full solution of both electron and hole transport equations, together with a suitable treatment of generation and recombination.

Drift-diffusion models

The most commonly used physical models are based on the drift-diffusion approximation and indeed all the early device models were based on this premise, obtained from the time-independent Boltzmann transport equation. This is a reasonable approximation for devices with relatively large feature sizes, where the carrier transit times far exceed the energy and momentum relaxation times. A relatively straightforward model emerges if the following assumptions are made: the gradient of the electron temperature ∇T_e is zero, the electron temperature remains close to equilibrium at the lattice temperature $T_e=T_0$. This leads to the well known *drift-diffusion approximation*, where the set of semiconductor equations reduces to the Poisson and continuity equations with the electron and hole current density expressions taking the form

$$J_n = qn\mu_n E + qD_n \nabla n \quad \textit{for electrons} \ , \tag{18}$$

$$J_p = qp\mu_p E - qD_p \nabla p \quad \textit{for holes} \ , \tag{19}$$

where n and p are the electron and hole densities, μ_n and μ_p are the electron and hole mobilities, D_n and D_p are the electron and hole diffusion coefficients. It should be noted that the current density equations are actually a simplified form of the momentum conservation equation, obtained from the second moment of the BTE. There has been considerable debate over whether the diffusion coefficient should appear inside the gradient operator. Note that the drift velocity of the electrons and holes is respectively

$$v_n = -\mu_n(E)E \ , \quad v_p = \mu_p(E)E \ . \tag{20}$$

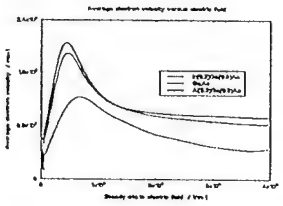


Figure 3. Velocity-field characteristics for electrons in GaAs, InGaAs and AlGaAs obtained from Monte-Carlo simulations.

The carrier mobilities μ_n and μ_p here are functions of local electric field, rather than average electron energy, and are often determined from the steady-state velocity-electric field characteristic. The steady-state mobility and diffusion coefficients are also a function of lattice temperature and doping level (Freeman and Hobson, 1972; Snowden *et al.*, 1983). The velocity-field characteristics for GaAs, InGaAs, and AlGaAs are shown in Fig. 3 (obtained from Monte Carlo bulk simulations). The diffusion coefficients are usually obtained from the Einstein relationships

$$D_n = \frac{kT}{q}\mu_n \ , \quad D_p = \frac{kT}{q}\mu_p \ . \tag{21}$$

Many semiconductor materials are anisotropic in nature (for example GaAs, and InP) which leads to tensor mobility and diffusion coefficients which depart significantly from the Einstein relations (Bauhann *et al.*, 1973; Snowden, 1982; Feng and Hintz, 1988).

Drift-diffusion models have been applied successfully to many semiconductor devices and many commercial simulation tools are based on this approximation. They are very well suited to modelling devices where the carriers do not experience significant carrier heating in the conducting regions of the device or where the transit time of carriers substantially exceeds the momentum and energy relaxation times. This is the case in most bipolar devices (diodes, BJT's, thyristors, triacs etc) and in many silicon devices (where the energy and momentum relaxation times are very short anyway). Hence, very good agreement has been obtained between measured and simulated device structures for MOSFETs with gate lengths of 1 micron or more and for most bipolar junction transistors. Difficulties arise when attempting to model many of the contemporary compound semiconductor microwave devices and some of the new generation silicon transistors. Here drift-diffusion models are often applied without due regard for their limitations and protagonists of this approach argue that the discrepancies can be eliminated by modifying the mobility and saturation velocities of the 'steady-state' velocity-field characteristics. Examination of the fundamental basis of the transport equations reveals this to be an unsatisfactory approach since there are other important features of hot electron transport that cannot be accounted for using these modified drift-diffusion models (for example hot-electron injection across interfaces and tunnelling). This situation illustrates the need for modellers to have a sound appreciation of the limitations of the models included in the software packages that are widely available today.

The majority of the closed-form analytical models available in the literature are based on drift-diffusion approximations. Hence, these should be applied with caution - even when using them as a means of obtaining simple estimates (simplistic modelling of short gate length MESFETs leads to large underestimates in the drain current and is a classic example of the gross errors which can occur using this approach).

Hot Electron Modelling

The non-stationary transport models derived from the Boltzmann Transport Equation are available in several degrees of approximation, varying from a full dynamic transport model to a simplified energy-transport model. The relative merits of the various approximations are discussed in detail elsewhere (Feng and Hintz, 1988; Snowden, 1988; Cook and Frey, 1982). This set of equations is often described as *hydrodynamic*. The description of carrier transport based on the drift-diffusion approximation assumes that the carrier energy distribution remains close to its equilibrium form. However, in many small-scale devices, such as microwave FETs, the electric field, carrier density gradient and current densities are often very large in magnitude. These high electric fields present lead to substantial carrier heating with the carriers attaining very high energies relative to the equilibrium levels. The carriers experience non-equilibrium transport conditions and their velocity may transiently exceed the equilibrium value. In these circumstances it is common for electron velocities to reach values of up to five times the 'steady-state' velocity. A rigorous treatment of the transport in small-scale devices requires the solution of the carrier, momentum and energy conservation equations described earlier. Examples of this approach will be given later in this chapter. In the earlier days of device modelling this rigorous approach was

both numerically and computationally challenging and the significance of hot carrier transport was not always appreciated (or relevant in the very early larger scale devices). Today, there are circumstances where a useful compromise is required between the detailed solution of the transport equations and the simplicity of the drift-diffusion model.

The complete set of non-stationary transport equations is often simplified by neglecting the terms $v.\nabla v$, the kinetic energy term $\frac{1}{2}m^*v^2$ (compared with $3kT_e/2$), and the time-dependence of the average carrier velocity. Here momentum and energy conservation equations for electrons reduce to the following forms

$$\mathbf{v} = -\frac{\tau_m(w)}{m^*(w)}\left(q\mathbf{E} + \frac{2}{3}\nabla w + \frac{2w}{3n}\nabla n\right), \tag{22}$$

$$\frac{\partial w}{\partial t} = -q\mathbf{v}\bullet\mathbf{E} - \mathbf{v}\bullet\nabla w - \frac{2}{3n}\nabla(nw\mathbf{v}) - \frac{w-w_0}{\tau_w(w)}. \tag{23}$$

This set of transport equations is often referred to as an *energy transport model*. Although most of the significant factors affecting the carrier dynamics are retained in this simplified model, there are some significant differences in the simulation results between the full transport model and the simplified form (Feng and Hintz, 1988). Further simplification is possible and two popular methods assume either a quasi-static approach (where $\partial w/\partial t$ and $\partial \mathbf{v}/\partial t \rightarrow 0$) or assume that spatial variation in energy and momentum are small (∇w and $\nabla \mathbf{v} \rightarrow 0$). In the first case, the momentum equation appears as in (23) and the energy conservation equation simplifies, as follows,

$$\mathbf{v}\bullet\nabla\mathbf{v} = -\frac{3}{5}\left(q\mathbf{E} + \frac{w-w_0}{\tau_w(w)}\right). \tag{24}$$

In the latter case the momentum and energy transport equations reduce to the form

$$\frac{d[m^*(w)\mathbf{v}]}{dt} = -q\mathbf{E} - \frac{m^*(w)\mathbf{v}}{\tau_m(w)}, \tag{25a}$$

$$\frac{dw}{dt} = -q\mathbf{E}\bullet\mathbf{v} - \frac{w-w_0}{\tau_w(w)}. \tag{25b}$$

A one-dimensional form of (24) was used by Carnez *et al.*, (1980) and Cook and Frey (1982) in the quasi-two-dimensional simulation of Si and GaAs MESFETs. The simplified forms of (25) have been used by a number of authors (Maloney and Frey, 1977; Shur, 1981).

The relative occupancy of the upper and lower valleys by electrons in multi-valley semiconductors can be taken in to account by modifying the momentum and energy conservation equations. Bozler and Alley (1980) develop a two valley model in terms of the energy separation between lower and upper valleys Δw_{UL} (0.36V for GaAs), the total energy w and the fraction of the electron population in the upper valley $G_U(w)$, where

$$G_U(w) = \frac{1}{\Delta w_{UL}}\left(w - \frac{3}{2}kT_e\right). \tag{26}$$

This expression is used to modify the energy transport equations as in Cook and Frey (1982) and Snowden and Pantoja (1989).

After establishing a suitable set of transport equations, the problem of simulating a device still requires further physical phenomena to be fully accounted for. Although early simulations were entirely based on relatively straightforward solutions of the basic transport equations, the agreement between measured and simulated data was often qualitative at best. In addition to requiring a set of boundary conditions which numerically satisfy the partial differential equations which represent carrier transport, simulations must also provide a realistic model of the contacts, surfaces and interfaces which bound the domain of the model. Furthermore, the description of the energy band-structure and material parameters in practical devices is often complex in nature. Finally, the impact of thermal effects on the operation of semiconductor devices is an important factor which is frequently omitted in models. The next section addresses physical processes and considerations which influence the accuracy and realism of the device model.

MATERIAL, GEOMETRICAL AND THERMAL PROPERTIES

The carrier transport model is only the foundation of any physical device model and the completeness of the analogy depends on providing suitable descriptions for the material, geometrical and thermal properties of the device in question. In some cases this can be a relatively simple step requiring straightforward boundary conditions and material parameters, as in the case of large diameter reverse-biased silicon Schottky varactor diodes. In contrast many important devices require a more detailed description of the material, device geometry and thermal considerations, such as in the case of microwave power transistors or sub-micron gate length digital MOSFETs. This section starts by considering generation-recombination and the role of traps in determining the behaviour of semiconductor devices.

Generation and Recombination

The significance of generation-recombination processes in any given device depends on the nature of the device in question. Generation-recombination plays a fundamental role in the operation of bipolar devices such as *p-n* junction diodes and bipolar junction transistors. In other devices, such as MOSFETs and MESFETs, generation-recombination has a negligible impact on the characteristics. However, even in the case of unipolar devices, generation-recombination mechanisms become important in the presence of high electric fields. The onset of breakdown in MOSFETs and MESFETs can occur at relatively low bias voltages and in these circumstances a suitable bipolar model which includes generation-recombination is necessary.

Generation-recombination processes are characterised as either direct or indirect in nature. The thermal generation of electron-hole pairs is an indirect process, attributable to phonon transitions occurring as a consequence of traps. In this process a trap centre in the energy band-gap determines the generation-recombination rate as a function of its capture and emission parameters. This generation-recombination mechanism is often represented using the thermal Shockley-Read-Hall model

$$G_{thermal} = \frac{n_i^2 - pn}{\tau_n(p+p_t) + \tau_p(n+n_t)}, \tag{27}$$

where τ_n and τ_p are the electron and hole lifetimes, n_i is the intrinsic carrier density, n_t and p_t are carrier densities dependent on the position and occupancy of the traps. The lifetimes τ_n and τ_p typically lie in the range 0.5 ns to 50 ms. Surface generation-recombination processes are often modelled using a modified form of the Shockley-Read-Hall model, replacing the carrier lifetimes with reciprocals of surface recombination velocities.

Impact ionization is a generation process, where an electron-hole pair is generated as result of interaction with a third particle. This process is only significant at relatively high electric fields where carriers attain high energies. The electron-hole pairs are generated by carriers moving directly across the band-gap. The impact ionization generation rate is given by

$$G_{impact} = \frac{1}{q}\left(\alpha_n |J_n| + \alpha_p |J_p|\right) \quad , \tag{28}$$

where α_n and α_p are the electron and hole ionization coefficients, which are a function of the magnitude of the component of the electric field parallel to the direction of current flow. The dependence of ionization coefficient for electron on electric field takes the semi-empirical form

$$\alpha_n = \alpha_n^{\infty} \exp\left[-\left(\frac{E_n^{crit}}{|E|}\right)^{\beta}\right] . \tag{29}$$

A corresponding relationship describes the ionization coefficient for holes. Typical values of the coefficients for GaAs and Si are given in Table 1, collected from the literature. There is significant variation in the values for E^{crit} and α^{∞} although the values for β are generally taken as 1 for silicon and 2 for GaAs (for both electrons and holes).

Table 1 Typical values of coefficients for impact ionization for electrons and holes

Material	Electrons α_n^{∞} (cm^{-1})	Electrons β_n	Electrons E_n^{crit} (Vcm^{-1})	Holes α_p^{∞} (cm^{-1})	Holes β_p	Holes E_p^{crit} (Vcm^{-1})
Si	7×10^5 to 2.2×10^6	1	1.1×10^6 to 1.75×10^6	6×10^5 to 2×10^7	1	1.4×10^6 to 3.3×10^6
GaAs	3×10^5 to 2×10^6	1 to 2 (often 2)	6.8×10^5 to 2.1×10^6	2.2×10^5 to 1.3×10^6	1.75 to 2 (often 2)	6.5×10^5 to 2.1×10^6

Auger recombination is the reverse of impact ionization and involves the recombination of an electron-hole pair and the emission of energy to a third particle. This recombination process involves both direct and indirect recombination processes in the presence of traps. The Auger recombination rate is given by

$$G_{Auger} = (n_i^2 - pn)(nC_n + pC_p) \quad , \tag{30}$$

where C_n and C_p are the Auger coefficients for electrons and holes which typically 3×10^{-31} cm^6s^{-1} and 1×10^{-31} cm^6s^{-1} respectively at 300 K for silicon. Auger coefficients have a weak temperature dependence unlike impact ionization coefficients.

Optical generation-recombination occurs via two mechanisms and is a direct band-gap process. Optical generation of electrons occurs when are excited from the valence band to the conduction band by gaining energy from incident photons. Radiative recombination occurs when an electron loses energy which is emitted as a photon and the electron moves from the conduction band to the valence band. Optical generation-recombination is not significant in silicon although it can be an important mechanism for direct band-gap semiconductors with relatively narrow band-gaps, such as GaAs. The optical generation-recombination rate $G_{optical}$ can conveniently be represented as

$$G_{optical} = C_{opt}(n_i^2 - pn) \quad , \tag{31}$$

where C_{opt} is the optical capture rate.

The total generation-recombination rate G is simply the sum of the individual contributions

$$G = G_{thermal} + G_{surface} + G_{impact} + G_{auger} + G_{optical} \quad . \tag{32}$$

In thermal equilibrium, a dynamic balance exists between the generation and recombination of electrons and holes. This balance is governed by the equilibrium relationship

$$n_0 p_0 = n_i^2 \quad , \tag{33}$$

where n_0 and p_0 are the electron and hole concentrations at equilibrium.

Traps and Deep Levels

The presence of traps in semiconductor materials can have a profound effect on device characteristics, determining indirect band-gap generation-recombination processes. The filling of traps is dependent on their concentration, capture cross-section, position in the energy band-gap and on the carrier distribution in the throughout the device. Acceptor sites are ionized when the trap energy level is below the Fermi level, whilst donor centres are ionized when the trap level is above the energy level. The trap filling process can be modelled using a modified Shockley-Read-Hall model, with recombination through a single level. The steady-state trap filling f of a particular level is given by

$$f = \frac{\tau_{cp} n_0 + \tau_{cn} p_0}{\tau_{cp}(n + n_0) + \tau_{cn}(p + p_0)} \quad , \tag{34}$$

where τ_{cn} and τ_{cp} are the time constants for capture and emission of electrons and holes respectively (Son and Tang, 1989). The electron and hole concentrations n_0 and p_0 are defined when the Fermi level is at the same energy as the trap and are given by

$$n_0 = n_i \exp\left(\frac{E_T - E_i}{kT}\right) \qquad p_0 = \frac{n_i^2}{n_0} \tag{35}$$

If the electron concentration is significantly higher then the hole concentration, which is often the case in unipolar microwave devices, the trap filling is controlled predominantly by changes in the electron concentration. This allows the filling factor f to be simplified to the form

$$f = \frac{n}{n + n_0} \quad . \tag{36}$$

The filling factor has to be calculated for each trap-level in the model. The sign of the charge resulting from each trap is determined by its type. Donor traps are neutral when filled and positive when empty, and acceptor traps are negative when filled and neutral when empty. Four level trap models have been shown to satisfy most requirements (Snowden and Barton, 1990). The Poisson equation is modified to take account of trap filling and in the case of a four-level model takes the form

$$-\nabla^2\Psi = \frac{q}{\varepsilon_0\varepsilon_r}\left[N_D^+ - n + p - N_A^- + N_{Dd}(1-f_d) - N_{Ad}f_a + N_{Ds} - N_{As}\right] \quad , \tag{37}$$

where N_{Dd} and N_{Ad} are deep donors and acceptors, N_{Ds} and N_{As} are shallow donors and acceptors, f_d and f_a are the filling factors of the acceptor and donor traps respectively. In this model the shallow donors are assumed to be completely empty and the shallow acceptors are assumed to be completely filled. The ionized deep-donor density for a single trap level with energy level E_T under steady-state conditions in a majority carrier device can be obtained using the following relationship (neglecting holes)

$$N_T^+ = N_T \frac{N_C \exp[-(E_C - E_T)/kT]}{n + N_C \exp[-(E_C - E_T)/kT]} \quad , \tag{38}$$

where E_C is the conduction band energy level and N_c is the effective density of states of the conduction band.

In the case of transient simulations it is necessary to solve the rate equation for traps in addition to the other transport equations, where

$$\frac{\partial (N_T - N_T')}{\partial t} = C_n n N_T' - e_n(N_T - N_T') - C_p p(N_T - N_T') + e_p N_T' \quad . \tag{39}$$

N_T is the trap density, e_n and e_p are the electron and hole emission rates. The capture coefficients C_n and C_p are given by

$$C_n = \sigma_n v_{nth} \quad , \quad C_p = \sigma_p v_{pth} \quad . \tag{40}$$

σ_n , σ_p are the capture cross-sections and v_{nth} , v_{nth} are the thermal velocities for electrons and holes respectively. The emission coefficients are related to the capture coefficient by

$$e_n = C_n N_c \left(\frac{1}{g}\right) \exp\left[\frac{E_T - E_C}{kT}\right] \tag{41}$$

$$e_n = g C_n N_c \exp\left[\frac{E_V - E_T}{kT}\right] \quad . \tag{42}$$

g is the degeneracy factor of the deep donor. The emission rate and trap energy levels are all temperature dependent (Lo and Lee, 1991). The transient stimulation of traps leads to changes in the trap filling which have time constant in the range 10ns to 50ms, where τ_n and τ_p are the minority carrier lifetimes where t_n and t_p are the minority carrier lifetimes determined as

$$\tau_p = \tau_{cp} = \frac{1}{\sigma_p v_{th} N_T} \quad , \tag{43}$$

$$\tau_n = \tau_{cn} = \frac{1}{\sigma_n v_{nth} N_T} \quad . \tag{44}$$

Although the response of traps is fundamental to the operation of bipolar devices, trapping phenomena remain in unipolar compound semiconductor devices where deep level traps in the substrate play an important role in determining the transient performance and substrate current of devices (Barton and Snowden, 1990). Traps play a fundamental role in determining the semi-insulating behaviour of GaAs and InP substrates. Semi-insulating GaAs behaves as an intrinsic material as well as possessing high resistivity. Deep level donor and acceptor traps compensate residual shallow acceptors or donors respectively to create semi-insulating substrates. Deep level traps are due to both unwanted impurities such as carbon, usually associated with the EL2 level, and to intentional doping with chromium. The impact of bulk traps is particularly significant for ion-implanted devices, but the time-dependent behaviour of surface traps remains important for epitaxial devices. Gallium arsenide semi-insulating substrates are often characterised using only one or two deep level traps - the EL2 midgap trap and a shallower trap with an activation energy of approximately 0.39 eV. A donor level due to Cr or EL2 occurs at E_C–0.75eV in GaAs. The Fermi level in Cr doped semi-insulating GaAs may also be fixed by donor levels at 0.89 and 0.45 eV below the conduction band edge. The EL2 level in GaAs has been found to have an emission time constant that has a strong dependence on electric field and a wide range of capture cross-sections and densities have been reported for varying from 4×10^{-20} to 2×10^{-18} m^{-2}.

Trap densities in active layers grown using epitaxial technology have lower values than those found in annealed ion-implantation samples. MOVPE layers grown on high quality semi-insulating substrates have been found to have very low electron and hole trap densities below 10^{17} m^{-3} and 2×10^{18} m^{-3} respectively in the active layer. The EL2 level is the dominant trap mechanism found in epitaxial layers grown on GaAs substrates at 0.81 eV, and is frequently the only one significant enough to be identified (typically at levels of 10^{20} m^{-3}). The EL2 trap level is likely to be present in the active channel beneath the gate as well as in substrate and buffer regions. The EL3 level is also found in some epitaxial samples, usually at relatively low densities below 10^{21} m^{-3}.

Surface traps can also play an important role in determining the electrical characteristics (Heliodore et al 1988). In particular the presence of surface traps strongly influences the breakdown process in GaAs MESFETs (Barton and Ladbrooke, 1985; Mizuta *et al.*, 1987). These traps, usually treated as deep level traps reflecting their position in the energy-band gap of the semiconductor, can act as either electron or hole traps. The total response of the device is due to a complex interaction between carriers in the active channel and those associated with the traps. The observed differences between steady-state DC and pulsed *I-V* characteristics of microwave FETs and low frequency dispersion associated with g_m and g_d (Ladbrooke and Blight, 1987) are attributable to the dynamic behaviour of deep level traps in the bulk material and at the surface of these transistors, Fig. 4.

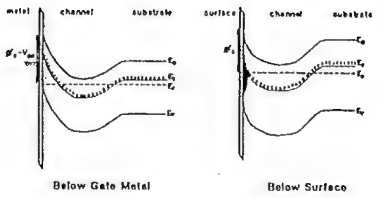


Figure 4. Equilibrium band diagrams below the gate metal and free-surface of a MESFET. Empty traps are donated by circles and full traps by filled circles.

The role of deep-level traps in the substrate of short gate-length MESFETs has been investigated using two-dimensional numerical simulations (Horio *et al.*, 1988; Barton and Snowden, 1990). This work has shown that substrate current becomes significant for FETs with substrates containing low acceptor and trap densities, such as is found in devices with thick epitaxial buffer layers. Zhao (1990) has modelled the effects of surface states on DLTS spectra of GaAs MESFETs, confirming that surface states do indeed exhibit hole-like DLTS signatures as observed by Blight *et al.*, (1986). This model includes deep-level traps in the active channel under the gated region and surface states on the ungated surfaces of the device.

Contacts

Ohmic contacts are usually modelled by simultaneously imposing charge neutrality and equilibrium conditions at the contacts, the boundary conditions for the source and drain contacts. This yields the following carrier density expressions:

$$n = \frac{1}{2}\left(\sqrt{(N_A - N_D)^2 + 4n_i^2} + (N_A - N_D)\right) , \tag{45}$$

$$p = \frac{1}{2}\left(\sqrt{(N_A - N_D)^2 + 4n_i^2} - (N_A - N_D)\right) . \tag{46}$$

The potential at the contact is modified by the built-in potential ϕ_b,

$$\Psi_{contact} = V_{applied} - \phi_b \ , \tag{47}$$

where the built-in potential of the contacts on n-type material is given by

$$\phi_b \approx \frac{kT}{q}\ln\left(\frac{N_D^+}{n_i}\right) \quad for \quad N_D^+ >> N_A^- \ . \tag{48}$$

A similar expression exists for contacts on p-type material.

Schottky barrier contacts can be modelled using varying degrees of approximation. In the case of reverse-biased contacts the simplest model sets the carrier concentration to zero and the potential to the applied voltage less the built-in potential. However, although this actually provides a reasonable representation in FETs operating with significant levels of reverse gate bias, it is not suitable for in many cases and does not take account of conduction current flow in the contact. A popular alternative model uses thermionic emission-diffusion theory, where the component of the current density normal to the contact J_n is specified in terms of the carrier density at the contact. Hence, in the case of n-type material, the electron density immediately below the contact n_s is given by

$$\mathbf{J}_n \bullet \mathbf{n} = qv_r(n_s - n_0) \ , \tag{49}$$

where v_r is the recombination velocity and n_s is the equilibrium electron density determined from

$$n_s = N_C \exp\left(-\frac{q\phi_{bi}}{kT}\right) , \tag{50}$$

where ϕ_{bi} is the built-in potential of the Schottky barrier (0.65 to 0.9 V).

Thermal Considerations

Carrier transport processes and material characteristics are strongly temperature dependent. Surprisingly, even though this is widely appreciated, the majority of device simulations assume that the entire device is at constant lattice temperature (often 300K - the ambient temperature). This is usually grossly inaccurate, with the exception of reverse-biased diodes, as many devices operate at elevated temperatures, with peak temperatures often 100K or more above ambient. This has a significant impact on parameters such as mobility, generation-recombination and trap occupancy, and in some circumstances the permittivity and other material parameters may be affected. Furthermore, the non-uniform nature of the temperature distribution contributes to the heat flow in the device, modifying the carrier transport process.

The heat flow equation describes the evolution of the temperature distribution within a device

$$c_L \rho_L \frac{\partial T_L}{\partial t} = \nabla \bullet (\kappa_L \nabla T_L) + H_S \ , \tag{51}$$

where c_L and ρ_L are the specific heat and density of the lattice (approximating the total heat capacity), κ_L is the lattice thermal conductivity, H_S is the heat generation and T_L is the lattice temperature. The heat flow equation couples with the carrier transport equations to form a coupled electro-thermal model. It is solved numerically, although the boundary conditions require careful consideration to ensure a satisfactory solution. If the thermal boundary is restricted to the same domain as the carrier transport equations, then equivalent third-order boundary conditions are necessary (Ghione *et al.*, 1987). Ghione *et al.*, (1988), who modelled MESFETs, have suggested that to obtain accurate results, the domain for analysis

should be extended horizontally for two to three times the source-drain contact spacing and to a depth of five to ten times the active layer thickness of the device. This represents an increase in the area of the simulation domain by typically a factor of six and has major implications for the choice of numerical method used to solve the transport model. The heat flow equation can alternatively be solved by dividing the analysis into an electro-thermal component over the original domain of the transport model and a non-heat generative model over a much larger region outside the transport simulation domain. This simplifies the boundary conditions and reduces the computational overhead.

The heat generation term is frequently evaluated from the scalar product of the electric field E and total conduction current density J,

$$H_S = \mathbf{J} \bullet \mathbf{E} + qE_g G \quad , \tag{52}$$

where E_g is the energy band-gap of the semiconductor and G is the generation-recombination rate.. The $J \cdot E$ product is Joule heating term, which describes the gain of carrier energy from the local electric field (Lisik 1981). The second-term on the right-hand side of this equation describes the energy exchange with the lattice through generation-recombination. This simple model of heat generation neglects band-gap narrowing effects.

The carrier mobilities display a strong temperature dependence which is reflected in the observed terminal current behaviour of devices. The temperature and electric field dependent mobility for electrons is often expressed in the following form:

$$\mu_n(E,T) = \frac{\mu_{n0}(T)}{\left[1+\left(\frac{\mu_{n0}(T)E}{v_{sat}}\right)^2\right]^{1/2}} \quad for\ Si$$

$$\mu_n(E,T) = \frac{\mu_{n0}(T) + v_{sat}\left(\frac{E^3}{E_0^4}\right)}{1+\left(\frac{E}{E_0}\right)^4} \quad for\ GaAs \tag{53}$$

where E_0 is the characteristic field for GaAs (2.69×10^7 Vcm^{-1}). The expressions for the hole mobility are similar to the silicon electron mobility relationship for both GaAs and Si (although the exponents in the denominator are both modified). The temperature-dependent low-field carrier mobilities follow a power-law dependence

$$\mu_{n0} = \mu_{n0}(300)\left(\frac{T_L}{300}\right)^{-\alpha_n}$$

$$\mu_{p0} = \mu_{p0}(300)\left(\frac{T_L}{300}\right)^{-\alpha_p} \tag{54}$$

where $\mu_{n0}(300)$ and $\mu_{p0}(300)$ are the low-field electron and hole mobilities at 300K. Typical values of the coefficients of these relationships are shown in Table 2. The temperature dependence of the saturation velocity of electrons is given by

$$v_{sat} = \left(\frac{T_L}{300}\right)^{-0.87} \times 10^7 \ cm/s \ for\ Si$$

$$v_{sat} = (1.28 - 0.0015T_L)\times 10^7 \ cm/s \ for\ GaAs \tag{55}$$

Two-dimensional thermal modelling of devices has recently been extended to thermal simulation of power monolithic microwave integrated circuits (Fan et al 1992). Fan et al used a control-volume finite-difference scheme which was solved using a Gauss-Seidel successive-over-relaxation scheme. They confirmed the presence of very large temperature gradients within the semiconductor material (GaAs).

Table 2. Coefficients for Si and GaAs temperature-dependent low field mobilities

Material	$\mu_{n0}(300)$ ($cm^2V^{-1}s^{-1}$)	α_n	$\mu_{p0}(300)$ ($cm^2V^{-1}s^{-1}$)	α_p
Si	1330 to 1600	2.2 to 2.6	440 to 600	2.2 to 2.7
GaAs	8000 to 9000	1 to 2.3	250 to 500	2.1 to 2.3

NUMERICAL SOLUTION METHODS

The solution of the set of coupled partial differential equations which constitute the transport equations generally requires the application of numerical methods. Although it is not within the scope of this chapter to provide a detailed treatment, it is worth commenting on some of the key aspects associated with the solution of classical transport model. The set of semiconductor equations chosen to represent the carrier transport process within a device requires discretizing over the domain of the model to obtain a set of non-linear equations which can be solved using either iterative or direct solution methods. In the case of one-dimensional models this is relatively straightforward process. The domain used to describe the region for numerical analysis of two-dimensional models is defined by the cross-section of the device in the plane of the current flow. Restrictions on computational resources limit the region of this domain, encompassing the active region of the device bounded by the edges of the ohmic contacts, the surface, and a suitable section of the substrate. This region is sub-divided into smaller sections to allow the semiconductor equations to be discretized across the domain of the model. The most popular numerical methods used to discretize these equations are the finite-difference, finite-element, finite-box and boundary element methods.

The solution of the discretized system of equations can be implemented using either a coupled (simultaneous) or decoupled (successive) schemes. Coupled solutions usually require an algorithm based on a Newton non-linear solution scheme (for example Kurata, 1982).The decoupled approach is often termed the Gummel Algorithm (Gummel, 1964). The reasons for choosing a particular method depend on the computing resources available and the type of simulation being performed. In circumstances where the mobile carrier densities exceed the fixed charge densities, the solution becomes highly non-linear and a coupled algorithm is generally the most more efficient approach. In simulations where the carrier densities are less than or equal to the fixed charge densities, decoupled solutions are more efficient. Gummel schemes decouple the equations and solve each equation

sequentially, iterating until the solution converges to the required accuracy. This is relatively to code, but it can be slow to converge, particularly in the case of bipolar devices where minority carrier levels require many iterations to reach the required solution (exceeding 1000 iterations in some cases). Nevertheless, this approach is particularly suited for use with simulations intended for computers with relatively small memory resources. The decoupled approach requires considerably less memory storage than the coupled Newton method when using direct solution methods. Time-domain two-dimensional unipolar simulations utilizing Gummel algorithms have been successfully written for use on personal computers with less than 1 megabyte of memory. Recently new decoupled algorithms for nonstationary transient simulations have been reported that allow time steps up to 50 times the dielectric relaxation time (Yoganathan and Banerjee, 1992).

Newton algorithms have quadratic convergence properties, which facilitate a rapid simultaneous solution of the transport equations. However, the computer memory storage requirements and relatively involved programming are considerably more demanding than for the Gummel algorithm. Drift-diffusion models which require the solution of the Poisson and continuity equations, are frequently implemented using coupled solution schemes, based on a modified Newton method. Comprehensive transport models, which incorporate energy and momentum conservation equations, are not as amenable to fast direct solution methods and sequential decoupled methods are more popular (Santos, 1991).

The numerical solution of the semiconductor equations is subject to stability and convergence problems. Numerical instabilities in the solution of the continuity equation usually manifest themselves as oscillations in the carrier densities. This may occur when using standard approaches such as linearized Crank-Nicholson central-difference schemes and the Galerkin finite-element method. These spurious oscillations can be suppressed in multi-dimensional simulations using modified techniques, such as upwind methods, although this may in turn result in numerical diffusion (Huang, 1985). These difficulties can be overcome by careful choice of algorithm (Shigyo *et al.*, 1989), with particular attention paid to the choice of mesh (Gresho and Lee, 1981). Discretization methods suitable for the energy and momentum conservation equations have been described by several authors (Rudan and Odeh, 1986; Snowden and Loret, 1987; Zhou and Ferry, 1992). In the case of the energy conservation equation it has been shown that the inclusion of the heat flux term (or a representation) significantly enhances the stability of the solution (Tomizawa, 1993). Generalised finite-difference solution schemes for the semiconductor equations have been described by Cole (1993).

The finite-difference and finite-element methods have been used extensively to solve the transport equations. Adaptive meshes with variable grid spacings are desirable in all two- and three-dimensional simulations to allow sufficient accuracy in regions of rapid change whilst conserving memory resources in areas of the domain where the variables are slowly varying. The finite-element method is probably the best choice in terms of efficient use of the mesh and computer memory resources and is particularly well suited for use with non-planar structures (for example recessed gate MESFETs, mesa-etched HBT's). However, the transport equations are generally easier to discretize using the finite-difference method which also has the advantage of extensively documented stability and convergence criteria. The following comments mainly apply to simulations based on finite-difference algorithms. The Poisson equation is readily solved using five-point central-difference discretization with either successive over-relaxation (SOR) or LU decomposition methods. In the case of Gauss-Seidel methods the SOR factor can be calculated analytically for rectangular regions (Varga, 1962) and values in the region of 1.8 are typically found to be close to optimum for most rectangular domains. There is no corresponding analytic method for determining the optimum relaxation factor for arbitrary non-rectangular simulation domains.

The current continuity equations are usually discretized using a half-point finite-difference notation for the current density terms. Explicit solutions of the time-domain continuity equations require prohibitively small time steps and these equations are often solved using the popular Scharfetter-Gummel scheme (Scharfetter and Gummel, 1969), which allows reasonable time steps. Alternatively, the $\nabla\cdot(n\mathbf{v})$ term in the continuity equation may be discretized using a second-upwind method, ensuring conservation in the discretized form (Roache, 1982). The time domain continuity equations utilize forward-difference schemes for the time-dependent term (even in finite-element implementations). Iterative Gummel simulation schemes require under-relaxation in the continuity equation algorithm because of the presence of complex eigenvalues in the Jacobi iteration matrix. This is often implemented using the Scharfetter-Gummel method in schemes based on electron and hole densities (for both finite-difference and finite-element methods). Straightforward semi-implicit Crank-Nicholson schemes are frequently used in time-domain continuity solutions, but timesteps are generally limited to very small values (often in the region of 10^{-14} s). Fully implicit schemes allow larger time-steps and have superior stability and convergence properties. Bipolar simulations often utilise quasi-Fermi level representations rather than explicit carrier densities in the continuity solution schemes. This allows the use of rapidly converging quasi-Newton scheme (Selberherr, 1984), although this can lead to errors in n and p because of the exponential dependence of the carrier densities on any error associated with the solutions for the quasi-Fermi levels. This necessitates the use of at least double-precision on most computer systems.

The energy conservation and momentum conservation equations are particularly susceptible to stability and convergence problems and require careful discretization. The $\mathbf{v}\cdot\nabla w$ term in the energy conservation equation can be discretized using a simple first-upwind scheme. The $\nabla(n\mathbf{v}w)$ term is conservative in nature and requires a second-upwind method or similar treatment to ensure a stable solution. The energy-conservation equation is often solved using a modified Scharfetter-Gummel scheme (Tang, 1984; Feng and Hintz, 1988). Central-difference schemes with half-point notation is used in the momentum conservation equation to discretize the $\nabla(n\mathbf{v})$. Feng and Hintz (1988) report that it is important to use a first-upwind discretization for the $\mathbf{v}\cdot\nabla\mathbf{v}$ to avoid numerical instability. The time-dependent terms in the energy and continuity equations can be discretized using backward-differences. Integration and expansion methods have also been used to address the time-domain solution (Bosch and Thim, 1974).

Boundary Conditions

The semiconductor equations require a set of consistent boundary conditions to obtain a solution for a given simulated structure (the 'domain'). This is usually achieved by assuming that the potential and carrier concentrations are fixed at a particular value (Dirichlet conditions) or have zero or constant derivatives (Neumann boundary conditions). Surfaces which exist between contacts or define the limit of the simulation inside the device structure are usually assumed to have Neumann boundary conditions (equivalent to establishing zero current flow across these internal boundaries). Alternative boundary conditions for the current-free surfaces may be formulated to account for surface and interface trapping effects, temperature gradients and passivation where there is a change in dielectric constant. The solution of the transport equations is by definition a strong function of the boundary conditions. It is equally important to appreciate that the accuracy of the numerical solution is sensitive to the implementation of the boundary conditions. The solution to the heat flow equation is particularly influenced by the accuracy of the numerical boundary condition (requiring a third-order scheme). Newton and Stirling polynomials satisfy most requirements for derivative boundary conditions (Snowden, 1988).

SIMULATION EXAMPLES

A very wide variety of established semiconductor devices have been modelled using traditional physical models. The choice of model depends on the nature of the device and its application. Large vertical devices such as junction diodes, where fringing effects at the edge of the mesa are negligible, can be satisfactorily represented using one-dimensional models. However, most devices require at least a two-dimensional model. In devices with interdigitated structures where the current flow is confined predominantly to a cross-sectional path between the contacts, a two-dimensional model will generally suffice. Although it is common to account for the passive three-dimensional nature of devices (contact resistances, parasitic static capacitances, scaling of contact currents due to finite widths and multiple cells), in practice there are few cases where three-dimensional solutions to the transport equations are required. Exceptions usually occur for small-scale devices where three-dimensional fringing effects are significant as in short gate width digital FETs (where the gate width is less than 10 microns). In many applications, which in principle require a two-dimensional model, it is possible to simplify the simulation, reducing the transport analysis to a quasi-two-dimensional case. This has been shown to be a powerful tool for studying FETs (Carnez *et al.*, 1980; Snowden and Pantoja, 1989,1992). This section will illustrate the application of two- and three-dimensional models and discuss the quasi-two-dimensional modelling concept.

Two-Dimensional Models

There are many examples of two-dimensional models available in the literature. The MESFET is an example of a device which has become a useful benchmark in modelling. The relatively simple structure and material composition of the MESFET is ideally suited to comparing different models. Feng and Hintz (1988) provide an interesting comparison of some of the most popular types of transport model for this type of transistor. MESFETs are essentially unipolar devices, usually fabricated from n-type gallium arsenide GaAs. The electron current flow in MESFETs is from the source to the drain, through a narrow conducting channel which is modulated by the gate depletion region under the control of the Schottky gate contact. In most practical MESFETs, with short gate lengths and small aspect-ratios (ratio of gate length to channel thickness), substrate injection plays a significant role in determining the current flow through the device. The doping profile in MESFETs varies over many orders of magnitude within a few hundred nanometres of the surface. These factors lead to a two-dimensional distribution of the potential, carrier densities, electric field and carrier energy. Two-dimensional FET models based on solutions of the carrier transport equations started to appear in the early 1970's, with the pioneering work of Kennedy and O'Brien (1970). In 1973, Reiser described a two-dimensional silicon MESFET simulation. This was followed by papers describing two-dimensional GaAs MESFET simulations (for example Yamaguchi *et al.*, 1976). The interest in short gate length devices for microwave applications stimulated work on modelling short channel GaAs MESFETs and hot electron effects (Wada and Frey, 1979; Carnez *et al.*, 1980; Curtice and Yun, 1981; Cook and Frey, 1982; Snowden, 1984). More detailed models began to appear later, solving both energy and momentum conservation equations (for example Snowden and Loret, 1987; Feng and Hintz, 1988). Recently simulations of short channel delta-doped MESFETs have been developed (Tian *et al.*, 1992). Lo and Lee (1992) have carried out numerical simulations of photoeffects in GaAs MESFETs using a two-dimensional model. They confirmed that the photovoltaic effect occurring at the channel/substrate interface is responsible for the substantial increase of the drain current when these devices are illuminated.

A two-dimensional electro-thermal simulation of a MESFET has been developed by Tsang-Ping *et al.* (1994) using a parallel implementation. This simulation solves an energy transport model togther with the heat flow equation for a relatively large simulation domain, Fig. 5. The semiconductor equations were solved using an SOR point iterative method with finite-difference discretization. The simulation was written for a parallel computer system using 16 transputers, which yielded a speed-up factor of 14.3 compared with a single processor.

Simulation results using the parallelized electro-thermal model are shown in Figs. 6-9 for a bias of V_{DS}=9 V, V_{GS}=–0.5 V. Note the relatively large domain of the simulation, encompassing large parts of the source and drain contacts and extending deeper in to the substrate than is usually included in isothermal simulations.

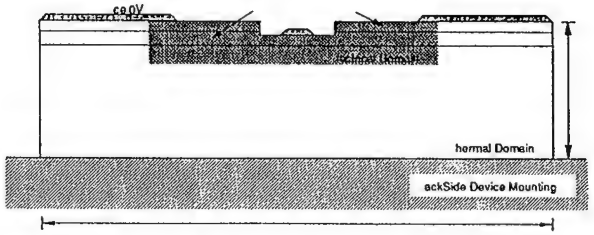


Figure 5 Two-dimensional domain for an electro-thermal simulation of a MESFET (Tsang-Ping *et al.*, 1994)

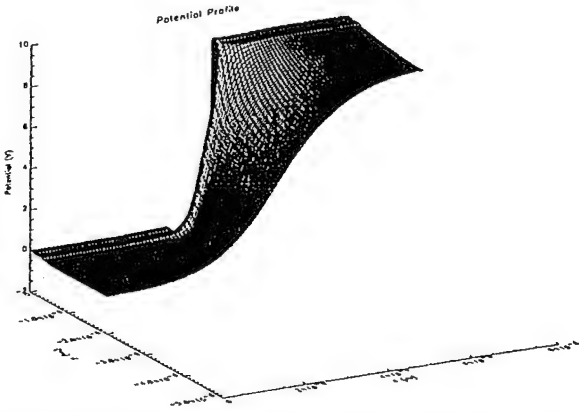


Figure 6. Potential distribution for the electro-therml model of the MESFET shown in Fig. 5 at V_{DS} = 9 V, V_{GS} = –0.5V.

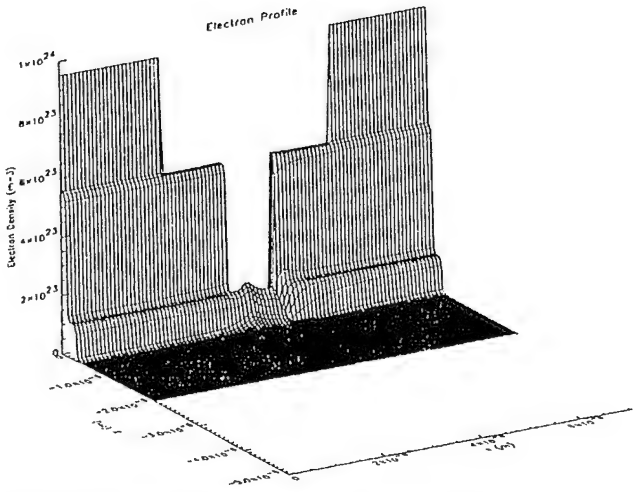


Figure 7. Electron distribution for the electro-thermal model of a MESFET shown in Fig. 5.

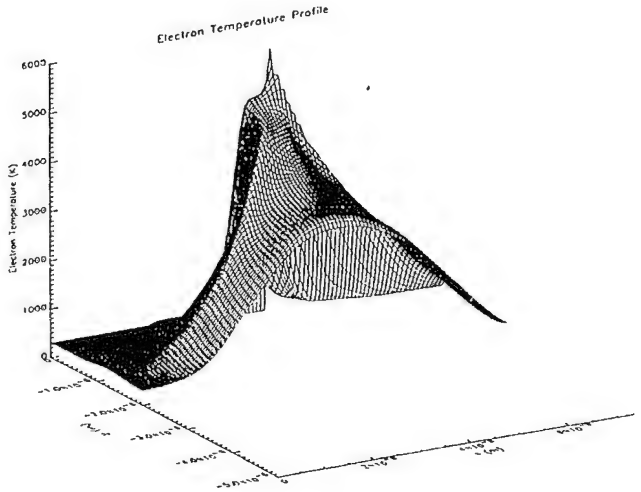


Figure 8. Electron temperature profile for the electro-thermal simulation of the MESFET in Fig. 5.

Drury (1993) has explored the simulation of HEMTs using two-dimensional semi-classical transport models coupled to a one-dimensional quantum mechanical description of the charge distribution in the heterostructure. This involes solving the current, energy and momentum conservation equations in two-dimensions taking account of the solution of the one-dimensional Shrödinger equation. This provides a more accurate model than the alternative classical quasi-Fermi level approach retaining the influnce of quanization in the two-dimensional electron gas region(s). A modified Newton scheme was used to solve the transport and charge-control model using finite-difference discretization. The simulation has been applied to AlGaAs/GaAs and pseudomorphic HEMTs. The simulation results for the 0.25 micron gate length AlGaAs/InGaAs/AlGaAs pseudomorphic HEMT of Figure 10 are shown in Figs. 11 to 13.

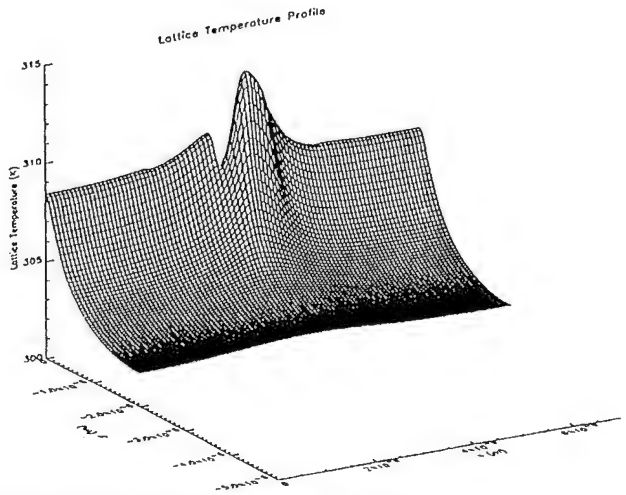


Figure 9. Lattice temperature profile for the electro-thermal MESFET simulation of Fig. 5

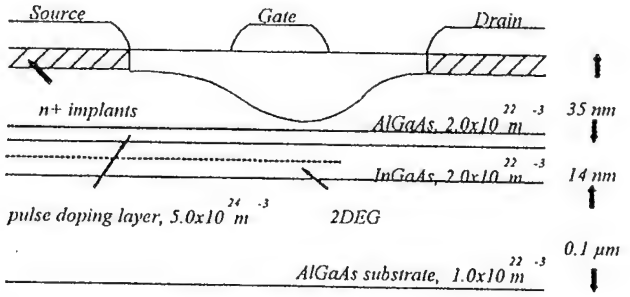


Figure 10. Two-dimensional cross-section of a *p*-HEMT.

An illustration of the impact of quantization on the modelling of *p*-HEMTs is shown in Fig. 14. Here the results from a classical (quasi-Fermi level) model are compared with the

results obtained using the classical/quantum mechanical model described above. It is evident that in this particular case the purely classical model overestimates the drain current. The differences between classical and quantum-hybrid models depend on the particular device. It has been found that for many AlGaAs/GaAs single channel HEMTs there is only a small difference between the results for quantum mechanical and quasi-Fermi level charge-control representations.

In some structures it is possible to use a quantum-moment approach which is analogous to the classical transport equations (taken as moments of the Boltzmann equation). This technique has been applied to MESFETs (Zhou and Ferry, 1992) and resonant tunneling structures (Grubin and Kreskovsky, 1989). The derivation and application of quantum models will be discssued elsewhere in this text.

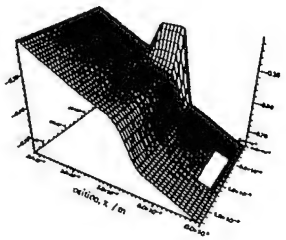


Figure 11. Fermi level profile for the two-dimensional *p*-HEMT model of Fig. 10.

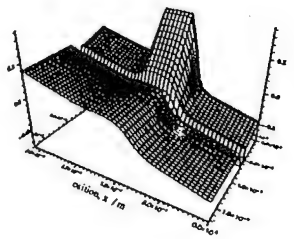


Figure 12. Conduction band profile for the two-dimensional *p*-HEMT model of Fig. 10.

Quasi-Two-Dimensional Models

Two-dimensional simulation of the semiconductor devices requires substantial computer resources and although this can normally be accommodated on modern workstations, it still requires a considerable amount of cpu time to simulate a particular device at a single bias-point. Today, there is an increasing demand for physical models to satisfy industry's demand for advanced design tools aimed at reducing the design time and cost of new devices (the 'right first time' philosophy). At present this is difficult to satisfy with two-dimensional simulations because of the prohibitive amount of time required to carry out the many simulation runs required to extract representative *I-V* characteristics, study effect of process variations and ultimately achieve a CAD perspective. The *quasi-two-dimensional approximation* overcomes the time constraints of the two-dimensional simulations. The concept of quasi-two-dimensional models has developed around FETs, although current research is applying this idea to other structures. Careful examination of full two-dimensional simulations reveals that the potential lines are almost parallel in the undepleted active channel and substrate of MESFETs and HEMTs, Fig. 15. This implies that the electric field and current flow in these regions is almost one-dimensional. The conducting channel is bounded by surface and gate depletion regions which can vary rapidly in short distances (hence this is not a gradual channel model!). Furthermore, the role of the substrate and active channel interface must be taken in to account.

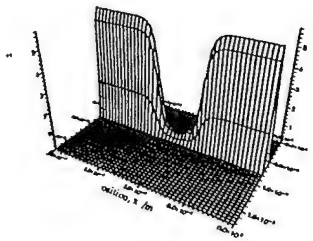


Figure 13. Electron density for the two-dimensional *p*-HEMT model of Fig. 10.

The first qualitative quasi-two-dimensional models appeared in the early 1980's (for example Carnez *et al.*, 1980; Cook and Frey, 1982). The accuracy of these models was improved by further development of the transport model and better estimation of the two-dimensional channel cross-section (Sandborn *et al.*, 1987; Darling, 1988; Snowden and Pantoja, 1989; Licqurish, 1989). A final stage of development in these models was to introduce a full charge-control model capable of taking full account of the doping profile, wafer structure, location and influence of traps and interface charge (Snowden and Pantoja, 1992). The quasi-two-dimensional model described by Snowden and Pantoja (1992) is intrinsically capable of simulating DC, small- and large-signal operation, without making any successive modifications to the model. The model has been used to *predict* the operation of new devices prior to fabrication. This makes the design of new devices systematic and

quantitative, minimising uncertainty in the design due to doping profile and fabrication process requirements. This type of model can be used to predict parameter spreads and yields prior to fabrication, since the device geometry and material parameters, which play a major role determining spreads and yields, form the basis of the model. A key feature of physical models is that it is possible to relate physical and geometrical changes in the device design to changes in the electrical and microwave performance. In particular, this model allows a wide range of design and process parameter variations to be investigated in a very short period of time without resorting to fabrication experiments.

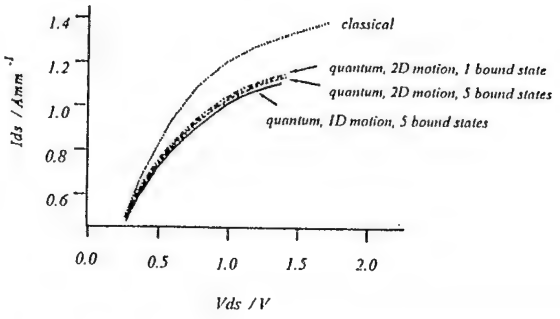


Figure 14. Comparison of *I-V* characteristics (V_{GS}=0) for classical models and classical models with quantum mechnical charge control, with one- and two-dimensional motion in the two-dimensional electron gas regions (Drury, 1993).

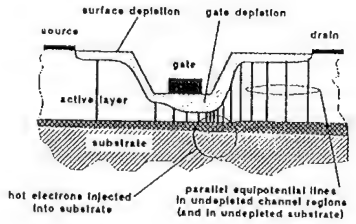


Figure 15. Quasi-two-dimensional FET model.

The quasi-two-dimensional model used in the work of Snowden and Pantoja (1992) incorporates a hot electron model solving both energy and momentum conservation equations. The model incorporates a charge-control model, doping profile description, a thermal dissipation model, surface and substrate trap representation and a flexible cross-sectional geometry description. The semiconductor transport equations are solved in a highly efficient quasi-two-dimensional manner (Snowden and Pantoja, 1989), retaining a two-dimensional description of the active channel, but only requiring a numerical solution of the *x*-component of the electric field parallel to the surface of the FET. The semiconductor equations for this model are Poisson's equation:

$$\frac{\partial E_Y}{\partial t} = \frac{q}{\varepsilon_r \varepsilon_0}\left[\int_{Y_D}(N_D - N_A + N_{DT} - N_{AT})dy - \int_{Y_D} n dy\right], \tag{56}$$

the current continuity equation:

$$i_S = i_{ch}(x) + i_g(x) + i_{aval}, \tag{57}$$

the momentum conservation equation:

$$v = \mu(w, T_0)\left[E - \frac{2}{3}K(w)\frac{dw}{dx} - \frac{k_1}{n}\frac{dn}{dx}\right], \tag{58}$$

and the energy conservation equation:

$$\frac{dw}{dx} = \frac{1}{1+\frac{2}{3}K(w)}\left[E - \frac{w - w_0}{v\tau_w(w)}\right], \tag{59}$$

where E is the electric field, w average electron energy, n electron density, v electron velocity, i_S, $i_{ch}(x)$, $i_g(x)$, $i_{aval}(x)$ are the source, channel, gate and avalanche multiplication currents, respectively, given by

$$\begin{aligned}
i_{ch}(x) &= ZY_{eq}\left[qn(x)v(x) + \varepsilon_0\varepsilon_r\frac{\partial E(x)}{\partial t}\right], \\
i_g(x) &= \left\{ZJ_0\int_z^{z+\Delta z} dx\left[\exp\left(-\frac{q[V(x) - V_G]}{kT_0}\right) - 1\right]\right\}\Bigg|_{L_{GS}<z<L_{GD}+L_G} + \frac{\partial Q_G(x)}{\partial t}, \\
J_0 &= A^* T_0^2 \exp\left(-qV_B / kT_0\right)
\end{aligned} \tag{60}$$

where Z is the gate width, Y_{eq} the effective undepleted channel height, $V(x)$ the channel potential, V_G the applied gate voltage, V_{Bi} the built-in voltage of the Schottky barrier, A^* is the effective Richardson constant (1.44×10^6 $Am^{-2}K^{-2}$), m the ideality factor (in the range 1.1 to 1.35 for the gate contacts of most FETs in the $1/kT$ region), T_0 is the lattice temperature. Y_D is the position of the limit of the Schottky depletion region calculated using either the approach of Sandborn *et al.* (1987) or Snowden and Pantoja (1989). K and $k_1(w)$ and are multi-valley transport parameters

$$K(w) \equiv \left[1 - \Delta_{LU} \frac{\partial G_U(w)}{\partial w}\right]$$
$$k_1 = \frac{2}{3}\left[w - G_U(w)\Delta_{LU}\right] \qquad (61)$$

It should be noted that the Poisson equation shown above fully accounts for both the doping profile (donor and acceptor) and for variable trap density where N_{TD} and N_{TA} are the ionized donor and acceptor trap densities. The Poisson, energy and momentum continuity equations are coupled to a simplified heat flow equation which yields the average temperature at each point in the channel, and solved using a finite-difference numerical scheme. The source current i_s and gate voltage v_G are used as a boundary conditions, from which the drain voltage v_{DS} is determined. The solution of these equations yields the terminal voltage and current responses.

The model, which already accounts for breakdown and gate conduction (Snowden and Pantoja,1989), allows both single and double-recessed FET structures appropriate for contemporary power FET designs to be simulated. The application of the thermal model requires an iterative solution, since the bias point is a function of temperature and the temperature itself is itself a function of bias (dissipated power). DC solutions using this thermal model typically require 4 or 5 iterations to converge. The quasi-two-dimensional model has the advantage of retaining a relatively rapid solution (typically 0.05 seconds per bias point on an IPX SUN workstation) compared with full two-dimensional models which are significantly slower.

The observed differences between steady-state DC and pulsed (fast transient) *I-V* characteristics and low frequency dispersion associated with g_m and g_d are attributable to the dynamic behaviour of deep level traps in the bulk material and at the surface of GaAs MESFETs. This quasi-two-dimensional simulation includes a charge-control model to determine the charge in the depletion region, charge in the channel and depletion depth as a function of channel to gate potential V_{CG} at each point in the channel. This allows the doping profile and influence of traps to be accounted for, achieving superior accuracy to that of earlier models. The Poisson equation is solved to obtain the potential V_{CG} with respect to the surface, as a function of depth through the profile. This involves determining the trap filling for each bias condition (channel-gate), to determine the ionized donor and acceptor densities. In order to simplify the solution, it is assumed that there is zero gate current in reverse-bias. The Poisson equation is solved by performing a numerical double integration of the total charge shown on the right-hand side, using a Simpson integration technique. This is performed for steady-state case, assuming that the traps reach equilibrium for each gate-channel bias and for a transient case (assuming $\tau < 1$ns), where the traps do not respond (V_{CG} is rapidly 'pulsed' from 0 V). This yields two separate the charge-control characteristics for steady-state DC simulations and for microwave simulations (including *S* parameters).

This physical model has been used to extract DC characteristics and time-domain microwave responses (S parameters and large-signal behaviour). The basic techniques for applying this model are described in Snowden and Pantoja (1989). The main purpose of this model is for use in CAD of devices and circuits which takes advantage of the relatively short execution times, typically less than 11 seconds for predicting a complete DC characterization at 150 bias points, on an IPX SUN workstation. Microwave S parameters typically require 9 seconds per frequency for a full c.w. time-domain and bias-dependent simulation on the same computer.

In order to determine the absolute accuracy of the model, a number of specific MESFETs were simulated, where the geometrical and physical parameters of the devices were carefully identified. This was done with the aid of cross-section SEM data, mobility profiling of material samples close to the devices, and careful process characterization. It is important to appreciate that all of the following results are obtained directly from the physical model, using the available physical data on the FET, without any additional fitting. A specific epitaxial 0.5 micron gate length MESFET was simulated under pulsed conditions to illustrate the transient capability of the model. In these circumstances it is assumed that the traps do not respond within the timescale of the pulse, and thermal dissipation is negligible. The simulated results were compared with pulsed wafer-probe measurements in Fig. 16. The agreement is good both in terms of absolute value and the slope of the characteristics, even down to pinch-off, confirming accurate modelling of the trap behaviour and the charge-control process. A specific 1 micron gate-length ion-implanted device, with a double implantation is used to produce an n^+-n profile was simulated over a range of DC bias conditions, (where the traps have attained equilibrium). The results for this device are shown in Fig. 17, where they are compared with wafer-probed DC measurements. Again the agreement is very good for this selected device. The correlation between simulated and measured data is good even at high values of $I_D.V_{DS}$ where self-heating becomes significan. This suggests that the relatively simple thermal model represents the heating effects in this device adequately.

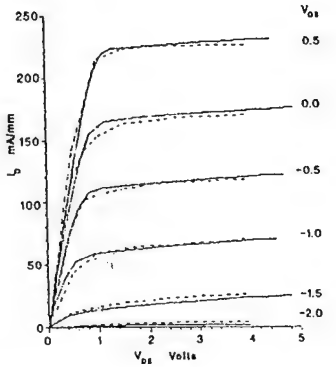


Figure 16. Pulsed *I-V* characteristics of a 0.5 micron gate length MBE MESFET. The solid line is simulated data and the dashed line is actual measured data (pulse duration <50 ns).

A comparison between simulated and measured *S* parameters for the 0.5 micron gate length FET is shown in Fig. 18. The measured data was obtained using wafer-probe techniques. The particular significance of these results is that they were obtained directly as a result of applying the physical model and process data, *without* any fitting to measured data or reverse-modelling. The model makes use of only the parasitics calculated during the simulation from the physical data, which includes the contact resistances, interelectrode capacitances and distributed nature of the gate.

The quasi-two-dimensional simulation approach has been successfully extended to modelling AlGaAs/GaAs HEMTs and pseudomorphic HEMTs (Veresegyhazy and Snowden, 1993). In this model a self-consistent solution of the Poisson and Schrödinger equations is used to obtain a charge control model for the layer structure. The Schrödinger equation takes the form

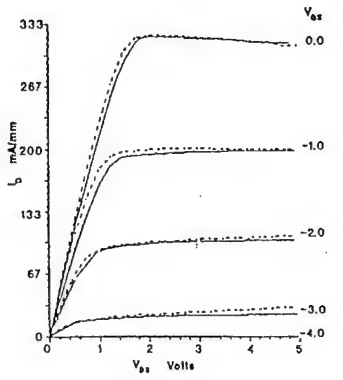


Figure 17. DC *I-V* characteristics of a 1 micron gate length MESFET. The solid line is the simulated results while the dashed line is the measured results.

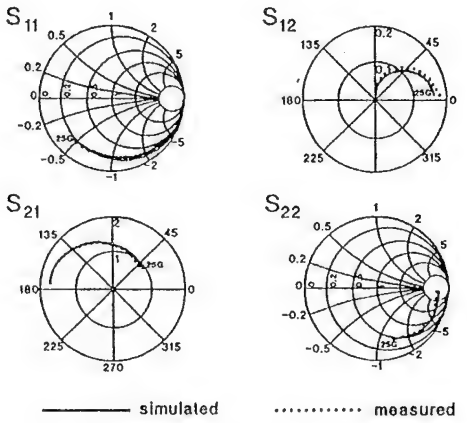


Figure 18. Comparison of microwave *S* parameters of a 0.5 micron gate length MBE MESFET.

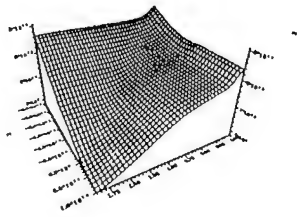


Figure 19. Two-dimensional charge-control model for *p*-HEMT using the quasi-two-dimensional of Drury and Snowden (1994)

$$-\frac{\hbar^2}{2}\frac{\partial}{\partial y}\left(\frac{1}{m^*}\frac{\partial \Psi_k}{\partial y}\right)+(U_E-\lambda_k)\Psi_k=0 \quad , \tag{62}$$

where Ψ_k is wave function corresponding to sub-band k, λ_k is the energy (Eigenvalue) at the bottom of sub-band k and U_E is the potential energy (in eV). The electron density n is then obtained from

$$n=\frac{m^* kT}{\pi\hbar^2}\sum_k\left|\Psi_k\right|^2 \ln\left[1+e^{-(\varphi_k/kT)}\right] \quad . \tag{63}$$

This accurately quantifies the sheet electron density attributable to the quantization in the two-dimensional electron gas layer(s) in these transistors.

Drury and Snowden (1994) have recently improved the model by including the influence of both components of electric field to obtain a Poisson equation of the form

$$\frac{\partial(\varepsilon_r E_x)}{\partial x}+\frac{\partial(\varepsilon_r E_y)}{\partial y}=\frac{q}{\varepsilon_0}(N_D^+-n-N_T^-) \quad , \tag{64}$$

where E_y is in the plane normal to the gate contact. The charge control model now requires the solution of the following equation

$$(y+y_s)\frac{\partial E_x}{\partial x}+\int dy\frac{\partial E_y}{\partial y}-\int N_D^+ dy+\int n dy=0 \quad . \tag{65}$$

Note that the transport equations still only require an essentially one-dimensional solution although they are a strong function of two-dimensions because of the coupling with the electric field. The two-dimensional charge-control model for a 0.25 micron *p*-HEMT (similar to the HEMT in Fig. 10) is shown in Fig. 19. A comparison of measured and simulated data for a commercial *p*-HEMT device is shown in Figs. 20 and 21.

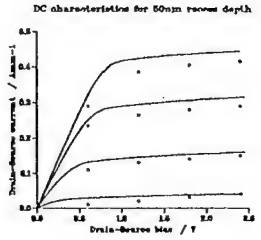


Figure 20. Comparison of measured and simulated DC charcteristics for a commercial *p*-HEMT.

Three-Dimensional Simulations

The use of three-dimensional semiconductor device simulations is usually restricted to very-small area devices. This is because most devices can be satisfactorily represented using two-dimensional cross sections coupled with straightforward scaling and because the computer resources required for the simulation of larger devices becomes prohibitive. There are essentially two approaches to three-dimensional modelling. The first approach utilises multipletwo-dimensional simulations by simulating `slices' of the device and associating each solution to extract a three-dimensional representation. This allows more modest computer facilities to be used together with existing two-dimensional codes.

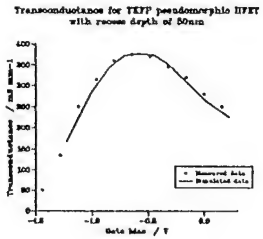


Figure 21. Comparison of simulated and measured transconductance for a commercial *p*-HEMT.

True three-dimensional simulations require extensive computer resources. Stability and convergence criteria impose limitations imposed on the cell dimensions which can be used. Hence, even if the variation in the third-dimension is relatively slow compared to that in the other two dimensions, the mesh refinement is constrained.

Buturla *et al.* (1981) simulated MOSFETs with 0.7 micron gate lengths and gate widths as small as 1.5 microns using the well known FIELDAY program. Their findings suggest that variations in channel length in these small devices may contribute more than 50% to the total threshold voltage tolerance of MOSFETs used in VLSI circuits. They found that a three-dimensional model was essential to accurately predict the source to substrate bias dependence of the threshold voltage. Further three-dimensional simulation of MOSFETs was carried out by Chamberlain and Husain (1981). Short and narrow MOSFET structures were studied by Salsburg *et al.* (1983) using a finite-element. Short channels were found to cause a decrease in threshold voltage because of the potential barrier lowering in the proximity of the high drain potential, whilst narrow channels have the opposite effect. McAndrew *et al.* (1988) have carried out three-dimensional simulations of MOSFETs including the momentum and energy conservation equations and accounting for heating effects and a three-dimensional variation in lattice temperature. They suggest that conventional two- and three-dimensional simulations based on the drift-diffusion approximation may underestimate diffusion currents. Limited computer resources and stability problems limited their simulations to low bias values.

Chan and Shich (1991) have reported a three-dimensional simulator for GaAs/AlGaAs heterojunction bipolar transistors. Their model is based on a drift-diffusion approximation implemented using a finite-difference scheme. This was solved using a Gummel algorithm and indirect sparse matrix solution techniques to minimise the computation time and memory requirements. In spite of these measures a Cray supercomputer was still required to implement the model and handle the matrices which had ranks of ten thousand or more! They used nonsymmetric preconditioned conjugate gradient package which provides an efficient vectorized sparse matrix solver. They showed that three-dimensional models are essential for devices with small geometry and investigated the effect of emitter grading on the current-voltage characteristics of an HBT 2×4 micron, with a 1×2 micron emitter. Another important feature which requires a three-dimensional representation is base current crowding in HBTs which do not have a self-aligned emitter.

Three-dimensional simulations are beginning to appear more frequently as computer resources improve. In particular the advent of relatively low cost parallel computation facilities has led to a number of new three-dimensional simulators (Brown *et al.*, 1993; Bodine *et al.*, 1993; Pennathur *et al.*, 1993).

CONCLUSIONS

Traditional classical and semi-classical models continue to satisfy the requirements for many device simulations. This chapter has outlined the basic principles of these models and illustrated their application to two- and three-dimensional device simulations. The advent of more powerful computers and the appearance of new parallel architectures has increased the interest in using physical models and stimulated the development of more sophisticated simulations. The drive towards ultra-small-scale structures will require more extensive use of quantum mechanical models and ultimately quantum transport simulations.

REFERENCES

Bauhann, P. E., Haddad, G. I., and Masnari, N. A., 1973, *Electronic Lett.*, 19:460.

Barton, T. M., and Ladbrooke, P. H., 1985, *IEEE Electron Dev. Lett.*, EDL-6:117.

Barton, T. M., and Snowden, C. M., 1990, *IEEE Trans. Electron Dev.*, 37:1409.

Blight, S. R., Wallis, R. H. and Thomas, H., 1986, *IEEE Trans. Electron Dev.*, 33:1447.

Bløtekjaer, K., 1970, *IEEE Trans. Electron Dev.*, 17:38.

Bodine, F., Holst, M. and Kerkhoven, T., 1993, *in* "Proc. Int. Workshop on Computational Electronics," Leeds, 70.

Bosch, R., and Thim, H. W., 1974, *IEEE Trans. Electron Dev.*, 21:16.

Bozler, C. O. and Alley, G. D., 1980, *IEEE Trans. Electron Dev.*, 27:1128.

Brown, A, Reid, D., Asenov, A, Barker, J., 1993, *in* "Proc. Int. Workshop on Computational Electronics," Leeds, 171.

Buturla, E. M., Cottrell, P. E., Grssman, B. M. and Salsburg, K. A., 1981, *IBM J.Res. Develop.*, 25:218.

Camez, B., Cappy, A., Kaszynski, A., Constant, E., and Salmer, G., 1980, *J. Appl. Physics*, 51:784.

Chamberlain, S. G. and Husain, A., 1981, *in* "Proc. IEEE Int. Electron Devices Meeting," IEEE Press, New York, 592.

Chan, H-C. and Shieh, T-J., *IEEE Trans. Electron Devices*, 38:2427.

Cole, E. A. B., 1993, *in* "Compound Semiconductor Device Modelling," Eds. C.M. Snowden, R.E. Miles, Springer-Verlag, London.

Cook, R. K. and Frey, J., 1982, *COMPEL*, 1:65.

Curtice, W, and Yun, Y-H, 1981, *IEEE Trans. Electron Dev.*, 28:954.

Darling, R. B., 1988, *IEEE Trans. Electron Dev.*, 35:.2302.

Licqurish, C., Howes, M. J., and Snowden, C. M., 1989, *IEEE Trans. MTT*, 37:1497.

Drury, R., 1993, "The Physical Modelling of heterojunction Field Effect Transistors", PhD Thesis, University of Leeds.

Drury, R., and Snowden, C. M., 1994, *in* " Proc. 3rd Int. Workshop on Computational Electronics, Portland.

Fan, M. S., Christou, A. and Pecht, M. G., 1992, *IEEE Trans. Electron Devices*, 39:1075.

Feng, Y-K., and Hintz, A., 1988, *IEEE Trans. Electron Devices*, 35:1419.

Freeman, K. R., and Hobson, G. S., 1972, *IEEE Trans. Electron Devices*, 19:62.

Ghione, G., Golzio, P., and Naldi, C., 1987, *in* "Proc. NASECODE V Conf.," Ed. J. J. H. Miller, (Boole Press, Dublin), 195.

Ghione, G., Golzio, P., and Naldi, C., 1988, *Alta Frequenza*, LVII:311.

Gresho, P. M., and Lee, R. L., 1981, *Comp. and Fluids*, 9:223.

Grubin, H. L., and Kreskovsky, J. P., 1989, *Sol.-State Electron.*, 32:1071.

Gummel, H. K., 1964, *IEEE Trans. Electron Devices*, 11:455.

Heliodore, F., Lefebvre, M., Salmer, G. and El-Sayed, O., 1988, *IEEE Trans. Electron Dev.*, 35:824.

Horio, K., Yanai, H. and Ikoma, T. 1988, *IEEE Trans. Electron Dev.*, 35:1778.

Huang, M. D. D., 1985, *IEEE Trans. Electron Dev.*, 32:2139.

Kennedy, D. P., and O'Brien, R. R, 1970, *IBM J. Res. Dev.*, 14:95.

Kurata, M., 1982, "Numerical Analysis for Semiconductor Devices," Lexington Books, Lexington, Toronto.

Ladbrooke, P. H., and Blight, S. R., 1987, *GEC J. Research*, 5:217.

Li, Q, and Dutton, R. W., 1991, *IEEE Trans. Electron dev.*, 38:1285.

Lisik, Z, 1981, *Sol.-State Electron.*, 24:85.

Lo, S-H., and Lee, C-P, 1991, *IEEE Trans. Electron Dev.*, 38:1693.

Lo, S-H., and Lee, C-P., 1992, *IEEE Trans. Electron Dev.*, 39:1564.

Look, D. C., Evans, K. R., and Stutz, C. E., 1991*IEEE Trans. Electron Dev.*, 38:1280.

McAndrew, C. C., Heasell, E. L., and Singhal, K., 1987, *Semicond. Sci. Technol.*, 2:643.

McAndrew, C. C., Heasell, E. L., and Singhal, K., 1988, *Semicond. Sci. Technol.*, 3:758.

Maloney, T. J., and Frey, J. 1977, *J. Appl. Physics*, 48:781.

Marashak, A. H., and van Vliet, K. M., 1984, *Proc. IEEE*, 72:148.

Mizuta, H., Yamaguchi, K., and Takahashi, 1987, *IEEE Trans. Electron Dev.*, 34:2027.

Pennathur, S., Ranawake, U. A., and Tripathi, V. K., 1993, *in* "Proc. Int. Workshop on Computational Electronics," Leeds, 156.

Pone, J. F., Castagnè, R. C., Courat, J. P. and Arnodo, C., 1982, *IEEE Trans. Electron Dev.*, 29:1244.

Reiser, M. 1973, *IEEE Trans. Electron Dev.*, 20:35.

Roache, P. J., 1982, "Computational Fluid Dynamics," Hermosa, Albuquerque, NM.

Rudan, M. and Odeh, F., 1986, *COMPEL*, 5:149.

Salsburg, K. A., Cottrell, P. E., and Buturla, E. M., 1983, "Process and Device Simulation for MOS-VLSI Circuits," Nijhoff, Amsterdam, 582-618.

Sandborn, P. A., East, J. R., and Haddad, G. I., 1987, *IEEE Trans.Electron Dev.*, 34:985.

Santos, J. C. A. D., 1991, "Modelling of Short-Gate-Length Metal Semiconductor Field-Effect Transistors for Power Amplifiers, PhD Thesis, University of Leeds.

Scharfetter, D. L., and Gummel, H. K., 1969, *IEEE Trans. Electron Dev.*, 16:64.

Selberherr, S., 1984, "Analysis and Simulation of Semiconductor Devices," Springer-Verlag, Vienna.

Shigyo, N., Wada, T., and Yasuda, S., 1989, *IEEE Trans. Computer-Aided Design*, 8:1046.

Shur, M., 1981, *IEEE Trans. Electron Dev.*, 28:1120.

Snowden, C. M., 1982, "Microwave FET Oscillator Development based on Large-Signal Characterisation", PhD Thesis, University of Leeds.

Snowden, C. M., Howes, M. J., and Morgan, D. V., 1983, *IEEE Trans. Electron Dev.*, 30:1817.

Snowden, C. M., 1984, *in* "Proc. Int. Conf. on Simulation of Semiconductor Devices and Processes," Pineridge, Swansea, UK, 406.

Snowden, C. M., and Loret, D., 1987, *IEEE Trans. Electron Dev.*, 34:212.

Snowden, C. M., 1988, "Semiconductor Device Modelling," Peter Peregrinus, London.

Snowden, C. M., and Miles, R. E., Eds., 1993, "Compound Semiconductor Device Modelling," Springer-Verlag, London.

Snowden, C. M., and Pantoja, R. R, 1989, *IEEE Trans. Electron Dev.*, 36:1564.

Snowden, C. M., and Pantoja, R. R., 1992, *IEEE Trans. Microwave Th. Techn.*, 40:1401.

Son, I., and Tang, T.-W., 1989, *IEEE Trans. Electron Dev.*, 36:632.

Tang, T.-W., 1984, *IEEE Trans. Electron Dev.*, 31:1912.

Tian, H., Kim, K. W., Littlejohn, M. A., and Bedair, S. M., 1992, *IEEE trans. Electron Dev.*, 39:1998.

Tomizawa, K., 1993, "Numerical Simulation of Submicron Semiconductor Devices," Artech House, Boston.

Tsang-Ping, C. S., Barry, D. M., and Snowden, C. M., 1994, *in* "Proc. 3rd Int. Workshop on Computational Electronics, Portland.

Veresegyhazy, R., and Snowden, C. M., 1993, *in* "Proc. Int. Workshop on Computational Electronics," Leeds, 96-100.

Wada, T., and Frey, J., 1979, *IEEE J. Sol.-State Circ.*, 14:398.

Wada, Y., and Tomizawa, M., 1988, *IEEE Trans. Electron Dev.*, 35:1765.

Varga, R. S., 1962, "Matrix Iterative Analysis," Englewood Cliffs, N.J., Prentice-Hall.

Yamaguchi, K., Asai, S., and Kodera, H., 1975*IEEE Trans. Electron Dev.*, 23:1283.

Yoganathan, S., and Banerjee, S. K., 1992, *IEEE Trans. Electron Dev.*, 39:1578.

Zhao, J. H., 1990, *IEEE Trans. Electron Dev.*, 37:1235.

Zhou, J.-R., and Ferry, D. K., 1992, *IEEE Trans. Electron Dev.*, 39:473.

MESOSCOPIC DEVICES - WHAT ARE THEY?

T. J. Thornton

Imperial College of Science, Technology and Medicine
Department of Electrical and Electronic Engineering
Exhibition Road
London, SW7 2BT UK

INTRODUCTION

It is hard to answer the question posed in the title of this Chapter because at present no real mesoscopic devices actually exist. But over the last few years there has been considerable exploration of the physics of mesoscopic structures with the hope that some of their physical properties might be exploited in devices of the future. Mesoscopic structures probe the properties of the solid state over energy and length scales not normally accessible to current day devices. We can conveniently group together mesoscopic devices into different classes within which the device action depends on the same physical length scale.

Two of the most important length scales are the phase coherence length and the mean free path. The former represents the average distance an electron will travel before its phase information is destroyed by an inelastic collision (typically 0.1 to > 1 μm) and is the relevant length scale for a class of mesoscopic devices known as quantum interference transistors. The average distance an electron travels before suffering an elastic collision is known as the mean free path. In the two dimensional electron gas (2DEG) of a high quality modulation doped GaAs:AlGaAs heterojunction the mean free path can exceed 10 μm and such a material can be used in another class of device where information is carried by beams of ballistic electrons.

Electron tunneling has played a central role in a variety of devices most notably the Esaki tunnel diode and the hot electron transistor. The distinction in the mesoscopic case is that the area across which tunneling occurs is reduced below 1 μm^2. At this length scale quantum confinement increases the energy of the electron states and modifies the I–V characteristics of the resonant tunneling.

Tunneling also plays an important (though not unique) role in Coulomb blockade around which single electron transistors (SETs) are based. The relevant energy scale here is the charging energy which is approximately $e^2/2C$. If the capacitance of the SET is small enough then $e^2/2C >> kT$ and no current will flow through the transistor until the applied bias exceeds the charging energy.

In what follows below the physics of four representative types of mesoscopic device structures are described along with their advantages and/or obvious disadvantages when compared to present day devices. Hopefully, these examples will help the reader to answer for his–or herself– 'what are mesoscopic devices?'.

QUANTUM INTERFERENCE TRANSISTORS

Interference of electron waves

Perhaps because of their many analogies with optical guided wave devices quantum interference transistors were among the first mesoscopic devices to be considered. In general the conductance of a quantum interference transistor is switched from the 'on' state to the 'off' state by changing the optical path length between two coherent electron waves such that the interference between them changes from destructive to constructive. When two electrons pass through a region of constructive interference a standing wave is established and the probability of finding an electron in this region is increased. In effect, the constructive interference has 'localised' the electron and if it occurs at the output or drain end of a quantum interference transistor its resistance will be increased. By way of illustration we will consider just the stub tuner interferometer although numerous examples of other quantum interference transistors have been discussed in the literature.

Stub Tuners

The mesoscopic stub tuner gets its name from the common practice in microwave engineering of inserting a needle or stub into a waveguide to minimise reflections from an unmatched load further down the guide. For the mesoscopic stub tuner Datta (1989) and Sols *et al.* (1989) have considered a geometry similar to that in Fig. 1. Electron waves traveling from the source can reach the drain directly or by the longer path via the gate, the difference in path length being DL. If the path length is reduced gradually, for example by increasing the reverse bias to the gate so that the edge depletion increases, the electron waves will undergo successive destructive and constructive interference each time DL is reduced by l_F the Fermi wavelength, and the conductance of the device will oscillate.

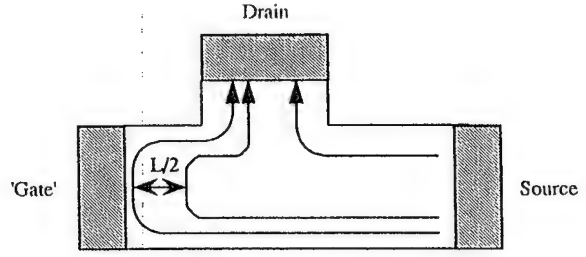


Figure 1. Schematic diagram of an electron wave stub tuner.

Datta (1989) has calculated the conductance as a function of depletion length L and the result, reproduced in Fig. 2(a), shows an almost 100% modulation in the conductance

between 'on' and 'off' states. Unfortunately, this ideal conductance modulation only occurs for the case of a single moded device ie one in which only a single 1D subband is occupied with a unique Fermi wavelength. For a multi–moded structure each occupied subband has its own Fermi wavelength and, in general, there is no optical path difference which each mode will simultaneously experience constructive interference. As a result the output characteristics are no longer nicely periodic and instead will consist of a number of Fourier components as shown in Fig. 2(b). This result shows the necessity of using single moded channels for any quantum interference device, a point we shall come across again for a different reason later.

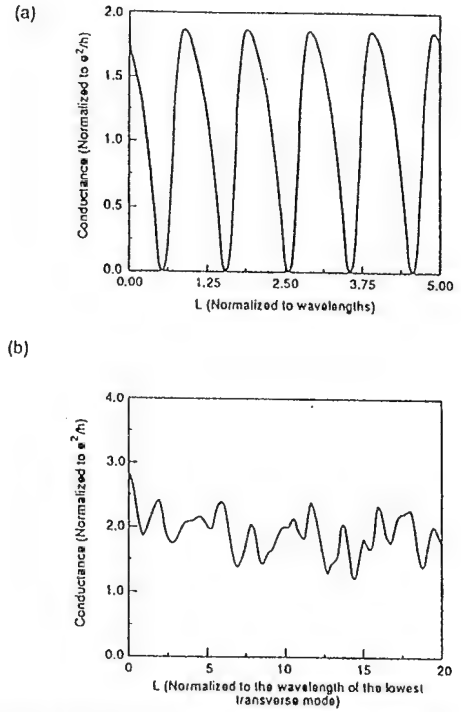


Figure 2. Numerical calculations showing transmission through the stub tuner of Fig. 1 for a) only one subband occupied and b) multiple subbands occupied.

Advantages of Quantum Interference Transistors

At first sight it might seem that the benefits of using an interference device arise from its small size. The intrinsic switching speed of a 0.1 μm long device with electron velocity of 10^5 ms^{-1} would be 10^{-12} s but these numbers are similar to those of a state–of–the–art GaAs MESFET or HEMT. In practice a quantum interference transistor would suffer from the same parasitic effects as conventional FETs and its switching speed would be limited, for example, by the RC charging time of the gate.

The main advantage of quantum interference devices compared to conventional FETs is that they are expected to have a very high transconductance, $g_m = dI_d/dV_g$, i.e. small changes in gate voltage lead to large changes in the drain current. In a MESFET or HEMT a comparatively large gate voltage is used to remove all the electrons from the channel by Schottky depletion. For the interference transistor the charge in the channel is essentially unchanged during the switching cycle, instead the optical path length is changed by $l_F/2$ with a comparatively small change in gate voltage. Unfortunately, for the one advantage of a high g_m the quantum interference transistor comes burdened with a a number of severe disadvantages.

PROBLEMS WITH QUANTUM INTERFERENCE TRANSISTORS

So far the biggest problem with quantum interference devices that we have come across is the requirement that only a single mode is occupied otherwise the conductance modulation will be less than 100%. To achieve single mode operation would require devices with cross–sectional areas of 10 nm x 10 nm. Assuming this size regime can be achieved in the not too distant future what other constraints will limit the applications of quantum interference transistors?

Low Temperature Operation

The main physical effect working against the practical applications of interference based devices is the fact that, at present, they can only function at low temperatures. At high temperatures the probability that an electron will have its phase memory randomised by (inelastic) phonon scattering increases and the phase coherence length drops rapidly. Ikoma *et al.* (1992) have measured the phase coherence length in GaAs:AlGaAs quantum wires and their results, reproduced in Fig. 3, show a rapid drop in t_ϕ ($L_\phi^2 \sim Dt_\phi$) for temperatures above about 5K.

If quantum interference transistors are to operate at higher temperatures some mechanism must be found to suppress the phonon scattering. Sakaki (1989) has suggested the use of superlattice quantum wires to suppress optic phonon scattering which is by far the most effective phase randomising process. In an earlier work Sakaki (1980) pointed out that elastic scattering in single mode wires is only possible between states at $+k_F$ and $-k_F$. Scattering events with such a large change in momentum, $\Delta p = 2\hbar k_F$, have low probability and as a result the electron mobility in a single mode wire is strongly enhanced. Unfortunately, optic phonons can scatter electrons between states of different energy but very similar momentum and will still limit the phase coherence length even in single mode wires. If instead of a uniform quantum wire a superlattice structure is used the continuous density of states for motion parallel to the wire breaks up into a number of minibands and minigaps. By choosing a suitable period for the superlattice it is possible to tailor the miniband structure such that the gap widths were greater than the optic phonon energy (36meV in GaAs) while the minibands themselves were narrower. As a result inter–band and intra–band scattering are strongly suppressed. Small energy scattering events are still of course allowed and are effective at randomising the phase but the suppression of optic phonon scattering should allow phase coherent transport over reasonable length scales at elevated temperatures.

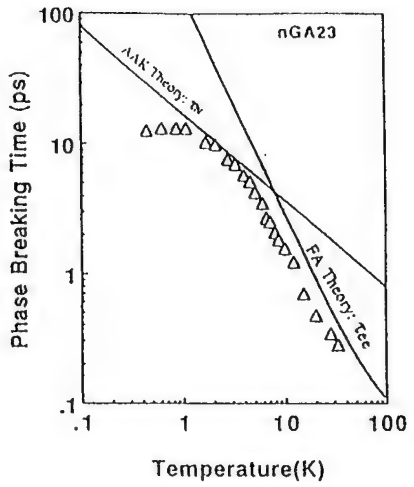


Figure 3. The phase breaking time measured as a function of temperature.

Random Impurities

All semiconductor materials include impurity atoms either introduced deliberately as dopants or incorporated unintentionally. In the comparatively large geometries of modern day devices there are so many impurities that their effect on electron transport can be considered in terms of their average properties. However, in a mesoscopic the active area can be as small as 0.1 × 0.1 μm and the effect of impurities must be considered on a microscopic level. In a GaAs sample an impurity concentration of 10^{15} cm^{-3} would be considered quite low but if this were used in the mesoscopic device considered above there would be at least one impurity atoms within 0.1 μm of the active area. The presence of one or two impurity atoms located at random sites will lead to significant differences in the electrical properties of nominally identical quantum interference transistors. This is because the dominant electron trajectories which determine the interference condition will now depend not so much on the device geometry but more by the impurity configuration. As a result the condition for constructive interference will differ from device to device and it would be almost impossible to define a threshold voltage. In effect each transistor would have to be individually characterised and supplied with its own data sheet.

High Resistance and Low Currents

In the 'on' state the resistance of a single mode quantum interference device will be >13 kΩ (see later). Although such a high resistance is undesirable from a circuits point of view it would be manageable if the current flowing through the device could be made arbitrarily large. Unfortunately, large currents lead to electron heating with a consequent reduction in the phase coherence length as described above.

To get some idea of the maximum current levels that might be achieved we can consider a GaAs quantum wire through which electrons travel ballistically. Such a wire would have the lowest possible resistance of 13 kΩ. Small applied biases would lead to incremental reductions in L_ϕ due to acoustic phonon emission. However as soon as the applied bias reached 36 mV the most energetic electrons will have enough energy to emit an optic phonon. Beyond this point all phase information will be lost before an electron can propagate the length of the wire and hence the maximum current we could drive through it and still maintain phase coherent operation is only I = 0.036V / 13 kΩ ~ 2.7 μA.

With such a small current drivability it would be hard to design circuits in which the output of one device can drive the inputs of several others. Of course the total current could be increased by connecting devices in parallel but the additional complexity of fabricating devices one on top of the other combined with their low temperature operation and inherent irreproducibility makes the future applications of quantum interference transistors really quite unlikely.

BALLISTIC TRANSMISSION DEVICES

For the quantum interference class of devices the relevant length scale was the phase coherence length which is typically a few microns at most. As mentioned in the introduction, the mean free path of electrons in a 2DEG can be much larger, often exceeding several tens of microns and electrons can propagate ballistically through the device ie with no collisions of any kind. Beams of electrons can be steered through the 2DEG using surface gates which behave like lenses and prisms. By switching the beams between different outputs we have a new class of ballistic transmission device. Before considering the refraction of electron beams we should first look at the quantisation of resistance and electron collimation that comes about when ballistic electrons propagate through a narrow constriction.

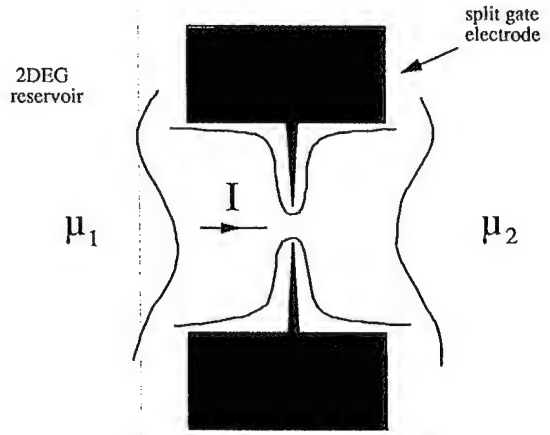


Figure 4. A split gate geometry suitable for the confinement of a quantum point contact. When a potential difference, μ_1-μ_2, is applied, current is forced to flow through the constriction

Electron Transport in Quantum Point Contacts

A quantum point contact (QPC) is a narrow constriction formed in a 2DEG by, for example, electrostatic depletion from a pair of split surface gates (Thornton *et al.*, 1986). A suitable geometry is shown in Fig. 4. A reverse bias applied to the gate electrodes will first deplete the underlying electrons but at a certain threshold voltage current can only flow from source to drain via a narrow constriction between the gate electrodes. The fringing field from the edge of the gate electrodes can be used to reduce the width of the constriction simply by increasing the reverse bias to the split gates. In a typical experiment the width of the constriction can be varied in the range 1.0 to 0.1 μm. The number of 1D subbands in the constriction is given roughly by the integer value of Wk_F/π and will therefore decrease as the width W of the constriction is reduced. The Landauer-Buttiker formula shows that the resistance of the point contact is given by $h/2Ne^2$ where N is the number of subbands in the constriction ($N \sim Wk_F/\pi$). If the resistance of the constriction is measured as a function of gate bias it will increase in a stepwise fashion as shown in Fig. 5. The plateaus are quantised to within a few per cent of the value given by $h/2Ne^2$ (Wharam *et al.*, 1988; Van *Wees et al.*, 1988).

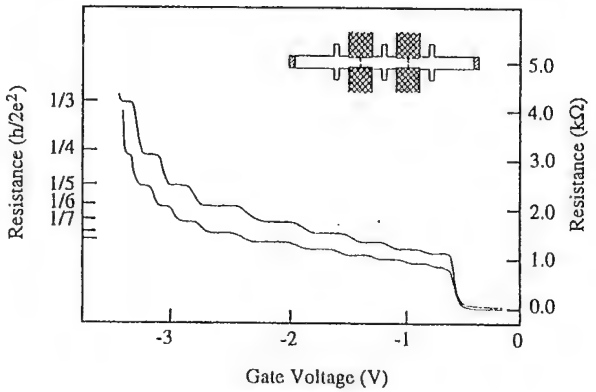


Figure 5. Quantized resistance plateaus in a QPC of variable width.

Another interesting property of a QPC is its ability to collimate a beam of electrons (Beenakker and Van Houten, 1989; Baranger and Stone, 1989). The phenomenon is well known in microwave engineering where a gently tapering horn can be used to collimate a microwave beam. A ray diagram illustrating the collimation is shown in Fig. 6. An electron entering the horn at some fairly large angle emerges with a trajectory which is more closely aligned to the axis of the horn. The collimation comes about because of the multiple reflections with the gently tapering walls of the horn. In a structure where the electron transport is ballistic the collimation can be maintained over fairly large distances and this property has been exploited in the electron refraction devices described below.

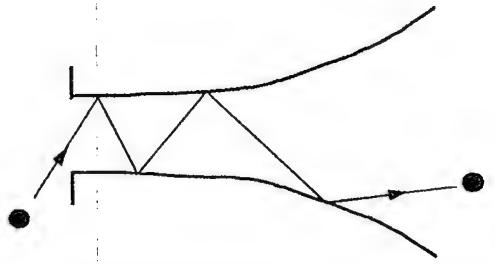


Figure 6. Electron collimation by a gently tapering horn.

Because of the reasonable accuracy of the resistance quantisation it has been suggested that the output characteristics of a QPC could be used to perform decimal logic or as elements within an analogue to digital convertor (ADC). Taking the last case as a specific example an analogue signal applied to the gate electrodes of the QPC would result in a 'digitized' output voltage if a constant current flowed through the QPC. The output signal could then be analysed using digital signal processing techniques, provided of course that the processing can deal with multiple voltage levels. The QPC analogue to digital converter would unfortunately only operate at low temperatures and is prone to the same irreproducibility as interference devices because the presence of an impurity atom within the vicinity of the constriction will alter the threshold voltage of the QPC and will disrupt the quantisation. Any advantage would come from the very small capacitance of the constriction but silicon based ADCs already operate at room temperature at frequencies of >1 GHz. Clearly silicon technology still has a lot of mileage left in it.

Electron Refraction in a 2DEG

Consider a 2DEG which has a gate deposited over half of the surface while the rest is left ungated as shown in Fig. 7(a). If the gate is reverse biased the electron concentration under the gate will be smaller than in the rest of the 2DEG. The conduction band will bend in a step-like fashion and there will be a force acting normal to the interface between gated and ungated regions. The parallel component of the momentum will therefore remain unchanged when an electron crosses the interface ie $p_1\sin\theta_1 = p_2\sin\theta_2$. The electrons which carry the current are within kT of the Fermi energy and we can write the momentum as $p=\hbar k_F$. For a 2DEG the Fermi wavevector is given by $k_F = (2\pi n)^{1/2}$ where n is the electron concentration (per unit area) and we obtain the result

$$\frac{\sin(\theta_1)}{\sin(\theta_2)} = \left(\frac{n_1}{n_2}\right)^{1/2} \quad (1)$$

The above equation was first derived by Spector *et al.* (1990a) and resembles Snells' law for the refraction of light at the interface between materials of different refractive indices. For the example of the partially gated 2DEG the trajectory of an electron crossing the interface will be deflected away from the normal when it moves from a region of high to low electron density (see Fig. 7b).

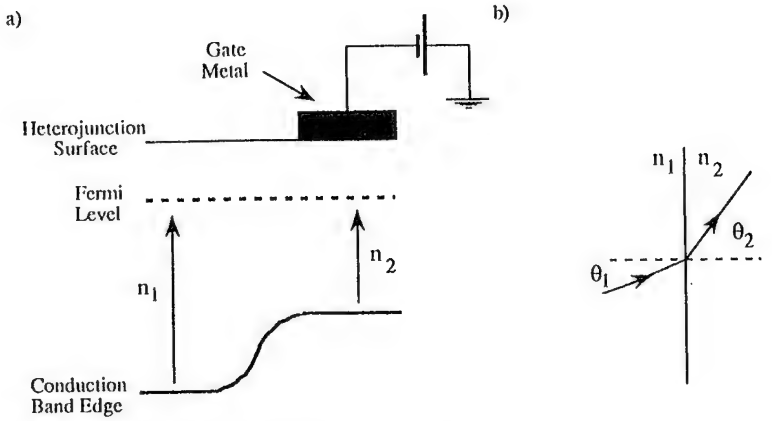


Figure 7. (a) At the interface between gated and ungated regions of a 2DEG there will be a discontinuity in the electron density (b) An electron crossing from a region of high to low electron density will be bent away from the normal.

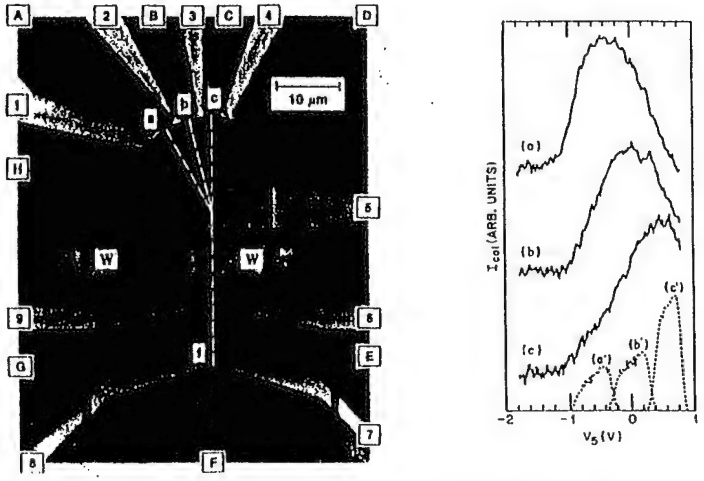


Figure 8. (a) The surface gate geometry used to refract a beam of ballistic. (b) The current flowing into each of the collectors at different gate biases.

Spector *et al.* (1990b) have used a triangular shaped surface gate to define a 'prism' in the underlying 2DEG. A plan view of their structure is shown in Fig. 8(a). The surface detail is quite complicated but the essential features of the device are the prism shaped gate, marked as [5], the injecting QPC marked as [*f*] and the three collector QPCs marked as [*a*], [*b*] and [*c*]. The distance between the QPC injector and collectors is smaller than the elastic mean free path so that electron motion is ballistic. With no bias applied to the prism gate we can assume that the electron concentration is the same everywhere (although in practice there is a reduction in density under the grounded gate) and a collimated beam of electrons injected from the lower QPC will propagate in a vertical direction and will be collected predominantly by collector *c*. However, as the gate is reverse biased the electron beam will be deflected and the majority of the current will flow into collector *b* and then collector *c*. The current flowing in each of the collector contacts as a function of the applied gate bias is shown in Fig. 8(b). Each collector shows a broad current peak at different gate biases confirming the basic operation of what might be called an electron refraction FET.

How would an electron refraction FET compare to the GaAs MESFET or HEMT? The answer is probably quite unfavourably if we are concerned only with the high speed performance which would be limited by the fairly large area and hence large capacitance of the prism gate used in the refraction device. However, the refraction FET is a good example of a transistor with the property of enhanced 'functionality' ie one which performs a function that would normally require several transistors. The enhanced functionality in this case is the possibility of switching multiple beams between different inputs and outputs. When a collimated beam passes through the 2DEG it doesn't interact with other beams which it may cross on the way and as a result information carried by the beams could be processed in a parallel fashion. Parallel signal processing is often associated with optical devices such as SEEDs (self electro-optic devices) which are themselves still very much in the research stage of development. The refraction FET would be limited to low temperature operation but unlike the quantum interference devices could probably be used at the much less expensive liquid nitrogen temperatures because the elastic mean free path is not as temperature sensitive as the phase coherence length. But to be realistic parallel signal processing using refraction FETS is unlikely to be explored seriously before optical processors are first developed.

RESONANT TUNNELING QUANTUM DOT DEVICES

The Esaki tunnel diode is an excellent example of a device with enhanced functionality. In its simplest form it consists of a heavily doped *p-n* junction with just two contacts and yet its *I-V* characteristic is extremely non-linear with the characteristic negative differential resistance (NDR) ie the current through the diode *decreases* even though the voltage across it is *increasing*. When such a device is suitably loaded it will act as a very high frequency signal source or amplifier and for this reason has received a considerable amount of attention. The latest devices are based on resonant tunneling through double barrier quantum wells and have been used in circuits operating above 400 GHz (Brown *et al.*, 1989).

One drawback of resonant tunneling diodes is that the bias at which the NDR occurs is determined by the geometry of the device and there is no mechanism for switching the NDR between different bias states. This kind of problem is generic to all two terminal devices and an obvious progression is to make a three terminal resonant tunneling device where the third electrode, which we shall call a gate in analogy with FETs, can be used to switch the NDR between 'on' and 'off states. In principle, the NDR due to resonant tunneling in quantum dot structures can be tuned by varying the cross-sectional area of the dot. Some of the different ways this has been done and the problems associated with small area tunnel structures are described below.

Resonant Tunneling Devices

The NDR that is so characteristic of the I-V curves of resonant tunneling devices has a similar origin to the NDR of the Esaki tunnel diode (Esaki, 1958). A simple 1D model can be explained using Fig. 9, which shows a generic double barrier resonant tunneling structure consisting of two heavily doped contacts separated on either side of an undoped quantum well by two thin tunnel barriers. The quantum confinement of electrons within the thin quantum well raises the energy of the first subband to a level $E_1 \sim \hbar^2\pi/2ma^2$ (where a is the well width) above the conduction band. For a narrow well E_1 lies above E_F at low bias and the resistance of the device is very high.

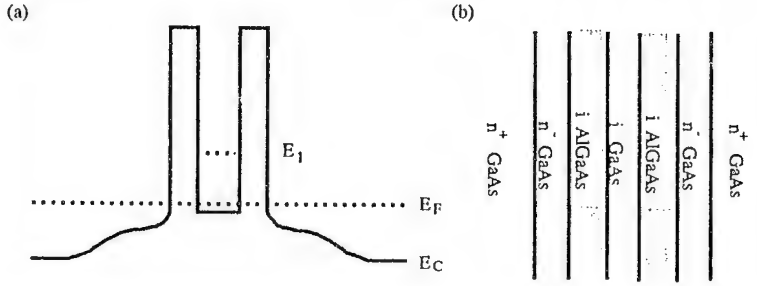


Figure 9. (a) Energy band profile of a double barrier resonant tunneling structure and (b) the corresponding layer structure.

The I-V curves of this simple structure can be understood using a model of sequential tunneling first proposed by Luryi (1985) which assumes that electrons can only tunnel if they conserve energy and momentum. Electrons in the emitter contact are free to move in all spatial directions and therefore have a 3D density of states in k-space occupying a Fermi sphere up to a radius $k_F = [2m^*(E_F - E_C)/\hbar^2]^{1/2}$. Electrons in the well are free to move in the x- and y–directions parallel to the well but the momentum in the perpendicular z-direction is fixed at a quantised value. To conserve energy and momentum the only electrons which can take part in the tunneling process are those contained within a circle formed by the intersection of the 3D Fermi sphere and a plane at $k_Z = q_W = [2m^*(E_1 - E_C)/\hbar^2]^{1/2}$ where E_C is the energy of the conduction band in the emitter (see Fig. 10a). If the transmission coefficient through the barrier is a constant (true at low bias) the current flowing through the device is proportional to the number of electrons which are free to tunnel which in turn is proportional to the area of the shaded circle in Fig. 10(b). By a simple application of Pythagoras' theorem we can now write that

$$I \propto \pi (k_F^2 - q_W^2) \propto \pi (E_F - E_1) \quad . \tag{2}$$

At vanishingly small bias $q_W > k_F$ and there is no intersection between the Fermi circle and the plane so that no current can flow. Relative to the conduction band in the emitter the energy of the state in the well, E_1, decreases linearly with bias and when $E_F - E_1 > 0$, the sphere and plane will intersect and current can begin to flow. The current will increase linearly with applied bias until E_1 drops below E_C, momentum perpendicular to the well can no longer be conserved and the current drops rapidly to zero.

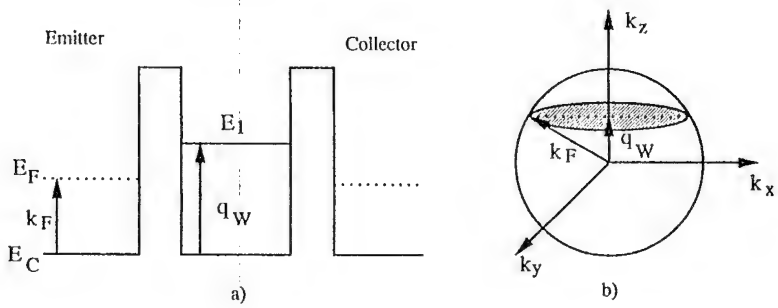


Figure 10. (a) Simplified energy diagram of a double barrier device and (b) a k-space diagram showing the electron states which can tunnel into the well.

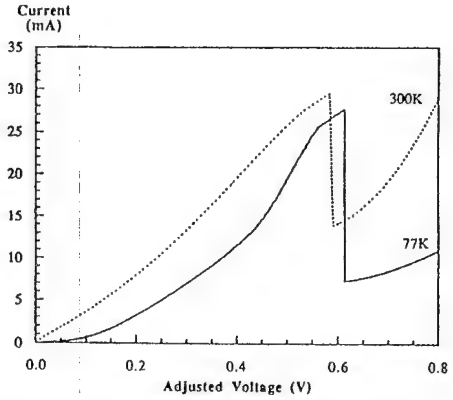


Figure 11. The I-V characteristics of a 10μm by 10 μm double barrier resonant tunneling structure at 300K and 77K.

Our conclusion from the simple model described above is that the current rises linearly when E_1 drops below E_F and falls to zero when E_1 drops below E_C. In a real resonant tunneling device the sharp NDR is superposed on a gently rising background (see Fig. 11). For biases below the NDR peak the current flowing can be small (at least at low temperature) and is due mainly to thermal emission over the emitter barrier. At higher bias thermal emission can contribute to the total current flowing but inelastic scattering processes also allow electrons to tunnel without conserving their energy. Inelastic tunneling destroys the coherence of the tunneling wavefunction and can be the dominant mechanism for current flow after the NDR peak has been reached. An important measure of the quality of a resonant tunneling device is the ratio of the maximum current at the NDR peak to the current minimum which comes afterwards, the so-called peak-to-valley ratio. Clearly, any

mechanism which increases the valley current will reduce the peak-to-valley ratio and degrade the performance of the device.

Resonant Tunneling in Quantum Dots

A drawback of the resonant tunneling device described above is that it is a 2-terminal structure. Once we have grown the double barrier layer by, for instance, MBE and then processed the wafer into a bonded device the I-V characteristic is fixed. It would be nice to have a third terminal to give the device a transistor action by which we could switch the NDR 'on' and 'off'. One way to do this is to pattern the quantum well layer into a quantum dot of variable width. A number of techniques have been developed to do just this but before discussing them in more detail we should consider how the extra quantisation present in a quantum dot structure effects the tunneling characteristics.

When the quantum well is confined in the x- and y- as well as the z-direction the single 2D subband of the wide quantum well splits into a number of higher lying quantum dot states. For convenience we can consider a quantum box with a square cross-section of side W. Assuming an infinite confining potential we can consider particle-in-a-box states and the energy of the lowest 2D quantum well subband will split into 1D states of energy E_n given by

$$E_n = \frac{\hbar^2\pi^2}{2m^*}\left[\frac{1}{a^2}+2\frac{n^2}{W^2}\right]. \qquad (3)$$

The new quantum dot energy levels are shown schematically in Fig. 12. When a bias is applied across the quantum dot the 0D states are swept passed the states in the emitter contact and assuming that the emitter is still fully 3-dimensional the simple argument outlined above would lead us to expect a number of NDR peaks in the I-V curves each one corresponding to a quantum dot state dropping below the conduction band of the emitter. By reducing the size, W, of the quantum dot we expect the energy and separation of the quantum dot states to increase. This in turn will effect the bias at which the NDR peaks occur in the I-V curve and in principle we can switch between different NDR peaks by changing the width of the quantum dot. An elegant way to vary the diameter of a quantum dot structure is to use depletion from surface or implanted gates. Both of these techniques can be used to reduce the conducting cross-section of the dot simply by applying a suitable bias to a third terminal as explained below.

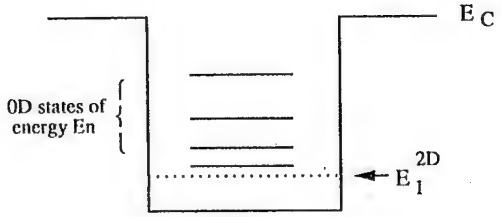


Figure 12. The formation of 0D states within a quantum dot.

Quantum Dots of Variable Diameter

Quantum dots with a fixed physical diameter have been made using reactive ion etching (Reed *et al.*, 1988; Tewordt *et al.*, 1990). The dots can be made as small as 0.1 μm with current being passed between a heavily doped substrate contact and a metal electrode deposited on a planarised surface. The fabrication of dots with a variable cross-section is significantly more complicated because of the need to place a gate electrode around the circumference of the ring without it shorting to the top contact. Usually, some kind of self-aligned is adopted whereby the top ohmic contact overhangs the partially etched pillar and behaves as a shadow mask for the gate electrode. The gating action can be achieved using surface gates or implanted junctions as shown schematically in Fig. 13. When reverse biased, the fringing field from the gate electrode reduces the conducting cross-section by Schottky depletion. For the case of the implanted gates, the *p*-type ring electrode circles the *n*-type dot and the width of the depletion region at the resulting *p-n* junction can be increased (decreased) by reverse (forward) biasing the gate implant (Goodings *et al.*, 1992).

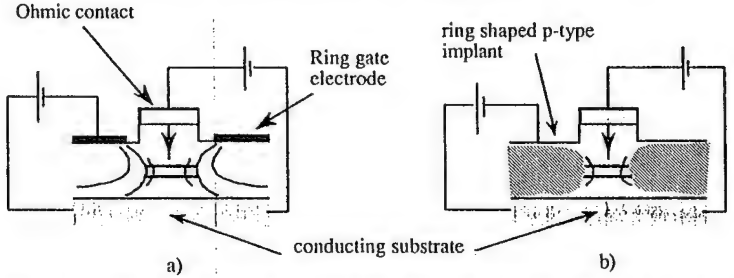


Figure 13. (a) Surface gating and (b) implantation approaches to the fabrication of variable area quantum dot structures.

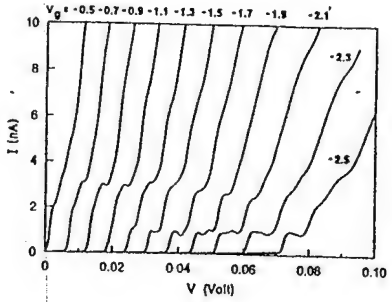


Figure 14. Experimental results of Gueret *et al.* (1992) showing multiple NDR peaks for a range of gate bias.

Both the Schottky and implanted gates can be used to vary the dot diameter in the range 0.1–1 μm. Results from a surface gated device (Gueret *et al.*, 1992) are reproduced in Fig. 14. These I-V characteristics do indeed show multiple NDR peaks which move to higher bias as the dot diameter is reduced. However, a simple interpretation based on the

simple model described above is fraught with difficulties, not the least being the assumption that the energy levels in the emitter and collector contacts do not change with gate bias. Not surprisingly, the ideal 3D-0D-3D geometry does not accurately model the shape of the quantum pillar which has a vertical cross-section which varies more like an hour-glass potential (Reed *et al.*, 1988). The current will flow by tunneling from 1D-like states in the emitter to 0D states in the well and fluctuations in the potential will lead to appreciable mixing of these states. Any attempt to understand the I-V characteristics of a real quantum dot tunnel device would require an accurate model of the confining potential and although this can be done in principle (Mizuta *et al.*, 1992) it is not a trivial task.

Another significant problem harks back to the random impurities that plague the quantum interference transistor. Although the barriers and well are nominally undoped there will be impurity atoms present primarily silicon atoms which segregate ie float on the surface during growth of the doped contacts. A background impurity concentration of 10^{15} cm^{-3} means that, on the average, there will be at least one impurity close to the tunnel barriers of even a 0.1 μm diameter quantum dot. The impurity atom will disturb the local crystal potential such that the barrier height and well depth will fluctuate randomly from device to device. Additional energy levels associated with the impurity atoms will be formed in the well and barriers and now resonant tunneling can take place not only via 0D levels but through the impurity atom as well. Additional NDR peaks will appear in the I-V curve at unpredictable bias values and the device is of little practical use because each one would have to be calibrated independently. Like the quantum interference transistor, the resonant tunneling quantum dot transistor displays a lot of interesting physics but any practical applications seem rather remote.

COULOMB BLOCKADE AND SINGLE ELECTRON TRANSISTORS

Single electron transistors (SETs) and other devices based on the Coulomb blockade rely on the fact that the energy to add an extra electron to a conductor body is proportional to the inverse of its capacitance. If its capacitance can be made very small then the charging energy is larger than the thermal energy kT or any applied bias eV and as a result no current can flow through the body. For an isolated conductor the capacitance will vary in proportion to its linear dimension and hence the smaller the object the smaller its capacitance. At the moment it is possible to make SETs with dimensions less than 0.1 μm and they operate at liquid helium temperatures but in principle we can wait for the technology to develop to the stage where we can make sub 0.01 μm devices which operate at or close to room temperature. For single electron devices, therefore, the relevant length scale is the linear size of the charging region and in general the smaller we make the device the better it will work. This is in marked contrast to conventional FETs where continued downscaling will mean that the MOSFET, for example, will cease to operate when electrons can tunnel through the gate oxide. At last it would seem that we have found a mesoscopic device which is not limited by some aggravating quirk of nature such as phonon scattering. Of course the reproducible fabrication of sub 0.01 μm structures is quite a significant challenge but at present SETs seem to have the brightest future of all the mesoscopic devices discussed so far.

Coulomb Blockade in Single Tunnel Junctions

Consider the current biased capacitor structure shown in Fig. 15 (Likharev, 1990). Although the conducting electrodes are separated from each other by a dielectric of some sort, the distance between them is small enough that at high bias an electron can tunnel from one electrode to the other. In essence this is a leaky capacitor or tunnel junction with a capacitance C and high field conductance G_T. The conductance in parallel with the junction G_S makes allowance for the fact that the environment around the junction eg the substrate, is not perfectly insulating.

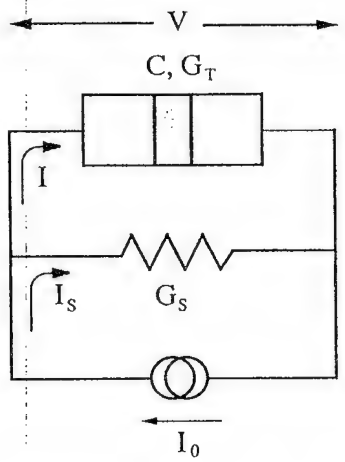


Figure 15. A current biased tunnel junction consisting of a thin barrier material (shaded region) sandwiched between two conducting electrodes.

In practice the area of the tunnel junction is made as small as possible to minimise its capacitance and it is reasonable to make the simplifying assumption that $G_S >> G_T$. Therefore as the current bias is slowly increased most of the current flows through G_S and the voltage across the junction is given by $V \sim I_S/G_S$. If current is to flow through the junction a charge of at least one extra electron must be stored on the capacitor and the increase in energy if this were to happen would be $e^2/2C$. The extra energy has to come from the potential driving the current bias and as a result no current can flow through the junction until $|eV| > e^2/2C$. The Coulomb blockade of current manifests itself in an ideal junction as a region of zero current flow for a voltage bias in the range $-e/2C < V < e/2C$ (see Fig. 16). For higher bias the junction begins to conduct and the conductance approaches a value given by G_T.

It has been predicted (Averin and Likharev, 1986) that a device with $G_T >> G_S$ will oscillate with a frequency $f_S = I/e$ when biased in the conducting branch of the I-V curve. The voltage oscillation is due to the correlated tunneling of single electrons but is very difficult to measure in practice. The reason is because of the large parasitic capacitance of the leads which have to be connected to the junction electrodes as well as the capacitance associated with the finite conductivity of the substrate. The parasitic capacitance increases the overall capacitance of the junction by so much that the charging energy is negligible compared to kT for any experimentally accessible temperature and as a result all effects due to the coulomb blockade are lost. Fortunately, there is a very effective means of decoupling a small conductor from the capacitance of its environment by using a double tunnel junction geometry.

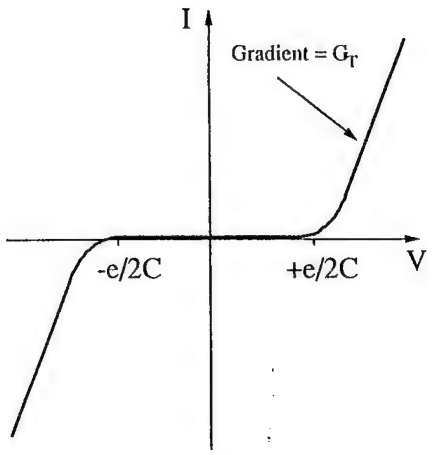


Figure 16. The *I-V* curves of an ideal tunneling junction at *T*=0 K.

Double Junctions and Single Electron Transistors

A circuit diagram for two small junctions connected in series is shown in Fig. 17(a) and a schematic picture of the geometry is given in Fig. 17(b). We are interested in the capacitance of the central conducting island which is simply the sum of the capacitances between the island and the left and right electrodes ie $C_T = C_L + C_R$. The effect of external parasitic capacitances are now less important unless they are coupled directly to the conducting island and of these the most important are the substrate and gate capacitances as we shall see later.

The energy required to add an extra electron to the central charging island of the double junction is ~ $e^2/2C_T$ and like the single junction case the *I-V* characteristics will display a region of coulomb blockade around zero bias. However, now the region of coulomb blockade is not necessarily symmetrically disposed around zero bias because in general the Fermi energy of the central island will not be the same as that in the leads. For a macroscopic device there would be a redistribution of charge between island and contacts until the Fermi energies were equal. But for the small capacitance conductor we are concerned with the electochemical potential can only be adjusted in discrete units of e/C_T. The *I-V* characteristics will still have a region of zero current flow over a voltage bias range of $\Delta V = e/C_T$ but it will be displaced from the origin by an amount $\Delta V_0 < e/2C_T$ as shown in Fig. 18. The voltage displacement can arise from different mechanisms and is hard if not impossible to predict. One possible origin is a difference in work function if the electrodes and charging island are made from eg different metals. A more important problem is due to polarisation charge induced on the island by the presence of random impurities close to the double junction. The impurities can be present in the substrate or in the tunnel barrier dielectric which is often a native oxide. The shift, ΔV_0, is essentially random and any circuits based on single electron transistors will have to make allowance for this.

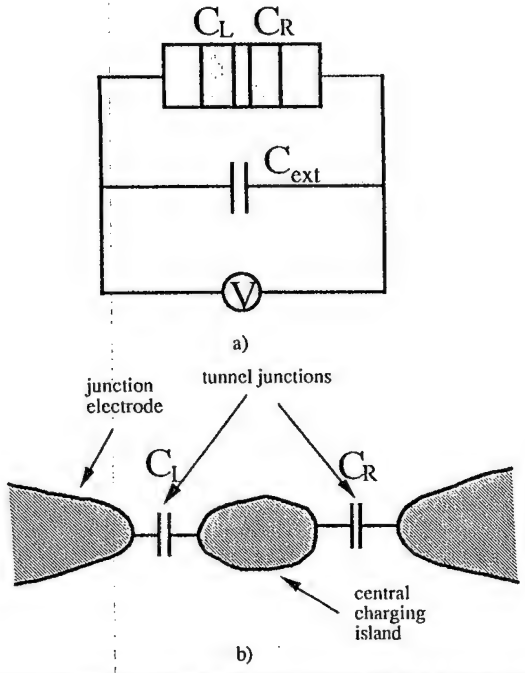


Figure 17. (a) Circuit diagram of a voltage biased double junction and (b) a schematic representation of a suitable geometry.

A single electron transistor can be made from the double junction geometry simply by adding a gate electrode as shown schematically in Fig. 19. The gate electrode is qualitatively different from the tunnel junctions in that, although it is capacitively coupled to the charging island, no electrons can tunnel through it. In practice this is done by placing the gate electrode far enough away from the island that the tunneling probability is negligible. The gate is used to alter the electrochemical potential of the electrons on the island by inducing a *quasi-charge*, $\Delta Q = CV_g$. Unlike the charge which tunnels onto the island via the leaky junctions the quasi-charge can be varied continuously because it is a polarisation charge.

We can use the concept of the quasi-charge to determine the switching condition of the single electron transistor. The energy of the charged island is still given by the simple expression $E = q^2/2C_T$ where q is the total charge on the island and C_T is the total capacitance given by $C_T = C_L + C_R + C_g$. If there are N excess electrons on the island the charge q is given by $q = (-Ne + \Delta Q) = (-Ne + C_gV_g)$ so that the total energy is

$$E = \frac{(-Ne + C_gV_g)^2}{2C_T} \quad . \tag{4}$$

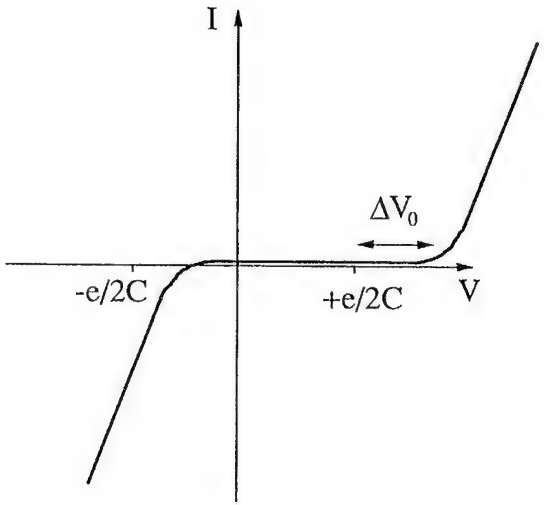


Figure 18. *I-V* characteristics showing the random shift in the Coulomb blockade in a voltage biased double junction.

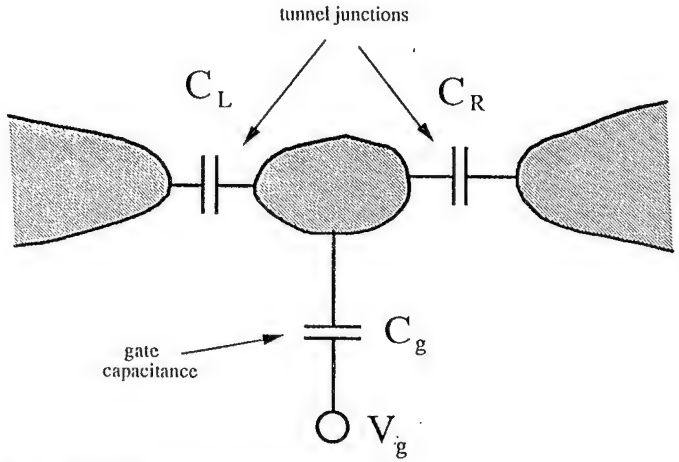


Figure 19. Schematic picture of a single electron transistor.

If a vanishingly small bias is applied across the double junction SET it will, in general, be in the regime of Coulomb blockade and no current can flow because the energy of the island with one extra electron will increase ie $E(N+1) > E(N)$. However, by adjusting the gate voltage we can achieve the condition that $E(N+1,V_g+\Delta V_g) = E(N, V_g)$ and now an electron can tunnel onto the island and a current will flow. By applying this condition and the above equation we see that an extra electron can tunnel onto the island each time the gate voltage is increased by an amount $\Delta V_g = e/C_g$ and in other words the SET will oscillate with increasing gate voltage with a period given by ΔV_g.

The Coulomb blockade oscillations have now been observed in a number of systems and results from Meirav et. al. (1990) taken from a split gate device are reproduced in Fig. 20. The oscillations are very clear and reproducible at low temperatures (T~50mK) but quickly disappear above temperatures of about 1K. The reason is due to the comparatively large capacitance introduced by the gate electrode. From the period of the conductance oscillations we can estimate a gate capacitance of $C_g = e/\Delta V_g = 8\times10^{-17}$ F for the smallest device. Assuming the gate capacitance dominates C_T we get an upper bound to the temperature at which we will still observe Coulomb oscillations of $T\sim e^2/2kC_g$. To achieve SET operation at room temperature is a major goal of much of the current research in this area which is inevitably focussed at reducing the gate capacitance. One method which does appear to give very small gate capacitance makes use of side gating in δ-doped layers.

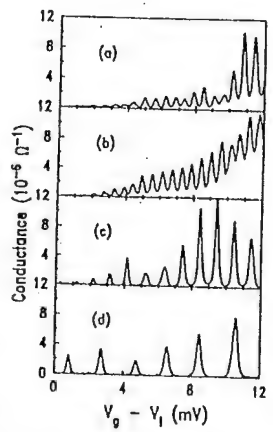


Figure 20. Coulomb blockade oscillations in a split gate SET.

Side Gated SETs, Multiple Tunnel Junctions and Co-Tunneling

Side gating, like the split gate technique, is a method for creating quantum wire structures of variable width. It was first applied by Wieck and Ploog (1990) to heterojunction wires and later developed by *Feng et al.* (1992) and Nakazato *et al.* (1992) to make δ-doped wires and constrictions. A scanning electron micrograph of a side gated constriction used to study single electron effects in δ-doped layers is shown in Fig. 21. A current bias is applied between source and drain electrodes on the left and right of the constriction. The side gate finger electrodes are separated from the δ-doped constriction by an air gap of ~0.5 μm. This structure has a threshold voltage of about -0.8 V and shows clear evidence of the Coulomb blockade as shown by the low temperature *I-V* characteristics in Fig. 22. The figure shows traces taken at different side-gate biases in the range $-0.7 \le V_g \le 0$ V. From the curve taken at 0 V the width of the Coulomb gap, $e/2C_T$, is ~5 mV giving a total capacitance C_T ~ 10 aF (1 aF= 10^{-18} F). As the gate voltage is reduced the zero field

conductance oscillates with a period of ~0.2 volts giving a gate capacitance of ~1 aF. These values are significantly smaller than those estimated for surface gated structures and side gated δ-doped SETs appear to be a useful system for developing high temperature single electron circuits.

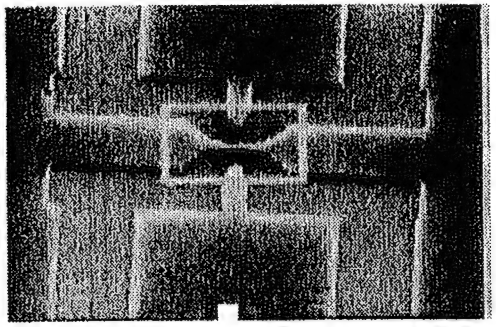


Figure 21. An SEM image of a side gated δ-doped single electron transistor.

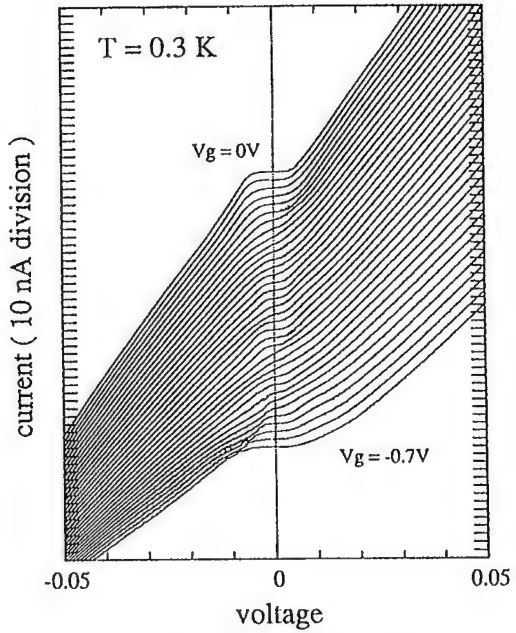


Figure 22. The I-V characteristics of a δ-doped SET for different gate biases.

A deeper analysis of the Coulomb oscillations from the δ-doped SETs shows there are several Fourier components present. This can be explained if we assume that the constriction has not split into one but a number of conducting islands connected in series. The exact number and shape of the conducting islands depends on the edge roughness and impurity profile along the constriction and is difficult to control in practice. Usually, only a small number (3-5) of islands are responsible for the observed effects and although the random nature of the δ-doped SETs is a nuisance the fact that they consist of multiple tunnel junctions has one very significant advantage, namely a dramatic reduction in the *co-tunneling* rate.

In the above discussion of the SET it has been implicitly assumed that the tunneling process through the device proceeds in a sequential fashion ie an electron first tunnels onto an island and remains there for some time before tunneling off again. The tunneling is not instantaneous but occurs at a rate given by (Averin and Likharev, 1992; Grabert, 1991)

$$\Gamma = \frac{1}{e^2 R_T} \frac{\Delta E}{1 - \exp\left(\frac{\Delta E}{kT}\right)}, \qquad (5)$$

where R_T is the resistance of the tunnel junction and ΔE is the change in energy of the system after tunneling ie $\Delta E \sim e^2/C_T$. At zero temperature this result gives $G = \Delta E/(e^2 R_T)$ and substituting for ΔE shows that the charging rate for a single junction device is governed by the charging time of the capacitor ie $G \sim 1/R_T C_T$.

It turns out that for the case of a double junction device an electron can tunnel through both junctions *simultaneously* via a process known as macroscopic quantum tunneling or co-tunneling. Any single electron circuit relies on the fact that an electron stays on a particular transistor until it is allowed to move on to the next and any unwanted tunneling events will reduce the reliability of the circuit. If an electron can co-tunnel through two or more junctions it can bypass one SET completely and upset the timing of the circuit. For a double junction SET with high resistance barriers the co-tunneling rate can be significant. However, if the SET uses multiple tunnel junctions as contacts to the charging island the co-tunneling rate can be dramatically suppressed and therefore the use of multiple junctions wherever possible is likely to increase the reliability of single electron circuits.

APPLICATIONS OF SINGLE ELECTRON TRANSISTORS

The beauty of single electron circuits is their potential for extreme miniaturisation and their low power dissipation. Power dissipation is a particular problem for the continued downscaling of silicon VLSI and by itself justifies further work on establishing circuit applications of SETs.

The single electron transistor is perhaps the ultimate logic device with the presence or absence of a single electron representing the digital states '1' and '0'. Likharev (1987) and Averin and Likharev (1992) have considered various elemental logic circuits and have come to the conclusion that, despite a relatively low voltage gain (Zimmerli, 1992), SET implemented logic is feasible. In this last part of the Chapter, rather than looking further at the integration of SET into logic circuits we shall concentrate on using SETs as current standards and as memory elements. The former exploits the unique ability of SETs to manipulate single electrons while the latter is an area which will benefit from low power dissipation.

The Single Electron Turnstile

Compared to the Ohm and the Volt which can both be defined with reference to physical phenomena (the quantum Hall and Josephson effects respectively) the Ampere is based on a less accurate standard. Turnstile devices are designed to clock single electrons around a circuit at an accurately defined frequency giving a current $I = ef$. If one and only one electron moves through the circuit each cycle then the current will be as accurately defined as the clock frequency and since timing can be measured extremely precisely, single electron turnstiles are expected to lead to significant improvement in the accuracy of the current standard.

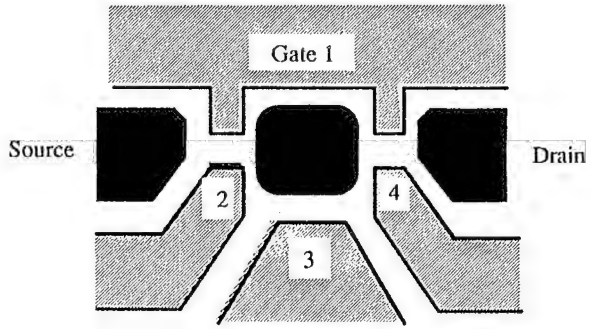


Figure 23. The surface gate arrangement of a single electron turnstile.

A single electron turnstile using oscillating tunnel barriers has been proposed by Odinstov (1991) and experimentally realised by Kouwenhoven *et al.*, (1991) who used a split gate structure to define the tunnel barriers. A schematic picture of the surface gate geometry is shown in Fig. 23. The gates are reverse biased to define the SET in the underlying 2DEG. The geometry is similar to the generic double junction shown in Fig. 17(b) with the exception that now the transmissivity (i.e. barrier resistance) of the tunnel junctions can be varied by means of the gate electrodes marked as 2 and 4. These electrodes are independently biased at microwave frequencies and modulate the barrier heights in anti-phase.

An energy profile across the device from source to drain is shown in Fig. 24. A sufficient voltage V_{SD} is applied to ensure that one, and only one, electron can tunnel onto the island during each cycle. As we have seen the tunnelling is not instantaneous and for the zero temperature approximation we can write a tunnelling rate $G = 1/R_T C_T$ where R_T and C_T are the resistance and capacitance of the tunnel junction. A high barrier will have a large resistance and the tunnelling rate will be small compared to the oscillating barrier frequency. When the barrier is lowered (Fig. 24b) the tunnelling probability increases and an electron can rapidly charge the island before the barrier is rained again (Fig. 24c) . Half the period later the right hand barrier is lowered allowing the electron to leave the turnstile and the process is ready to repeat giving an average current $I = ef$.

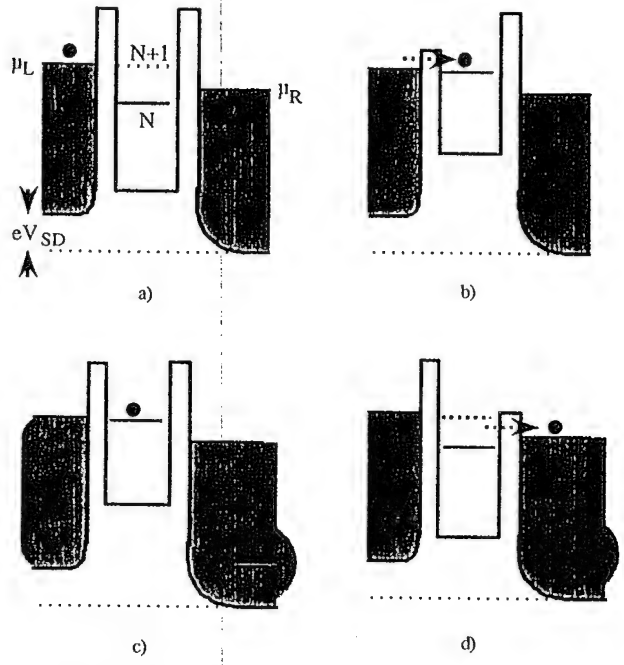


Figure 24. (a) to (d): The clock cycle required to drive a single electron through the turnstile.

The gate modulation frequency and the number of electrons flowing through the turnstile determine the current flowing. More than one electron could be clocked through each cycle and provided the number of electrons N remained constant the current would be quantise at a value of $I = Nef$. However, if electrons can pass through the turnstile randomly by; for instance co-tunnelling, then the accuracy of the current quantisation will be reduced. As we saw earlier, multiple tunnel junction reduce the co-tunnelling rate and will have to be used in turnstile devices if they are to provide current standards with an accuracy of better than 1 part in 10^7.

Single Electron Memory

Averin and Likharev (1992) have shown that under certain circumstances a capacitor charged by means of a single electron transistor can display hysteresis with potential memory applications. Consider the situation in Fig. 25(a) where a capacitor is being charged through a leaky tunnel junction. If there is an excess n, of electrons the energy of the system is given by

$$W(n) = \frac{ne^2}{2C_T} + neV\frac{C}{C_T}\ , \qquad (6)$$

where $C_T = C + C_0$. The first term in the equation for $W(n)$ represents the charging energy of the node M while the second term is the work done by the polarisation charge, neC_0, flowing through the battery in the external circuit.

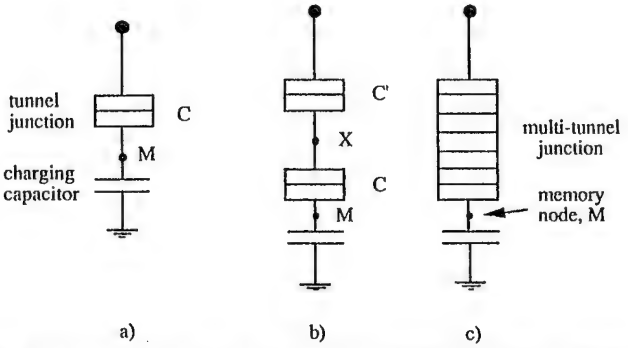


Figure 25. Charging a capacitor through a) a single junction, b) double junctions and c) multiple junctions.

If we consider the system in its ground state (n=0) we can add an electron to the node M by increasing V until $W(n) = W(1)$ and this occurs at value $V = -e/2C$. By changing the applied bias we can add or remove electrons from the node M but this circuit has no memory action -- as soon as the applied bias is removed the system will drop back to its ground state. The reason is because the energy expression for $W(n)$ is uniquely defined by the value of n. Averin and Likharev (1992) have pointed out that a bistable situation occurs for the modified circuit of Fig 25(b) where the capacitor is now charged through a series combination of two tunnel junctions. For the case $C' = C = C_0$ the energy is given by (Averin and Likharev, 1992)

$$W(n,n') = \frac{e^2}{3C_0}(n^2 + n'^2 - nn') + (n - n')\frac{eV}{3} \quad , \tag{7}$$

where n and n' are the number of electrons which have passed through the two capacitors. The double junction memory all exhibits bistability since we can choose different values n, n' and V which have the same energy. For instance, for an applied bias $V = -e/C_0$ the energy of the states $W(0,0)$ and $W(1,0)$ are equal ie an electron can tunnel onto the memory node as long as the charge at X is zero. On the other hand, $W(1,1) = W(0,1)$ at V=0 and an electron on the memory node can be removed provided an electron is present at X. For the latter case the charge at X induces a polarisation charge which partially cancels the charge at the memory node thereby compensating for the lack of applied voltage. As with other single electron circuits the memory cell is susceptible to co-tunnelling but this can be suppressed by using a multiple tunnel junction to charge the capacitor as shown in Fig 25(c).

A single electron memory cell operation at 4.2K has been demonstrated by Nakazato *et al.* (1993). They used side gated δ-doped SETs as described earlier to form tunable multiple tunnel junctions (MTJs). The equivalent circuit and an SEM image of the memory cell are shown in Fig. 26 and Fig. 27. The single electron memory consists of two separate circuits, the actual memory cell itself and an electrometer to measure the voltage at the memory node. The capacitance of the electrometer should be small enough that it doesn't significantly increase the total capacitance of the memory node. A gated multiple tunnel junction SET is therefore a good choice. The electrometer is biased by a voltage V_{cg} such that it is close to the threshold condition ie very nearly turned off. In this condition it behaves much like an FET close to threshold. The memory node is capacitively coupled to the electrometer and any change in its potential will lead to an increase or decrease in the current flowing through the electrometer. During the experiments the current through the electrometer varies almost linearly with gate voltage and is therefore a good indicator as to the voltage on the memory node.

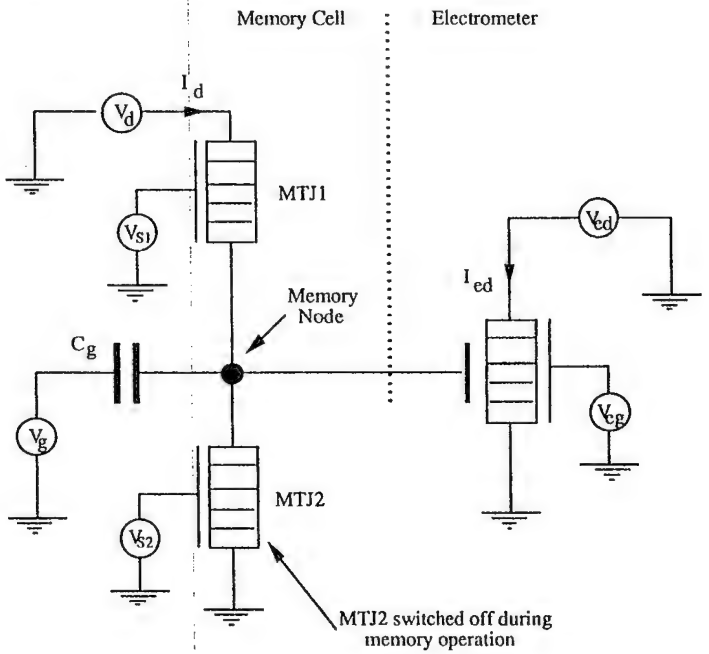


Figure 26. The equivalent circuit of the single electron memory cell.

The memory cell itself resembles the circuit of Fig. 25(c) except now two MTJ SETs are connected to the memory node (MTJ1 and MTJ2). The two SETs are useful for characterisation of the memory cell and during operation the memory node is only charged through one SET, the other being switched off by applying a large reverse bias to the side gate.

Fig. 28 shows how the current through the electrometer, I_{ed}, varies as the gate voltage V_g is cycled and illustrates the pronounced hysteresis which gives this circuit its memory action. Consider increasing V_g so that we move from A to B. No current can flow through MTJ1 which is in the regime of Coulomb blockade. The applied voltage V_g is dropped across the series combination of C_g and C_{MJTI} which together behave as a potential divide and the voltage at the memory node (and therefore I_{ed} in the electron) varies linearly with V_g.

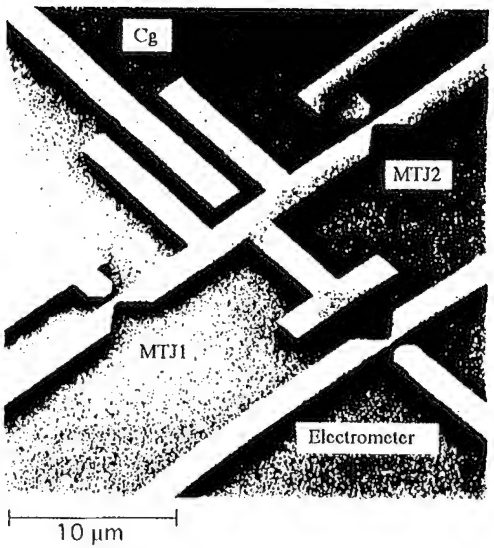


Figure 27. Scanning electron image of the surface of the single electron memory.

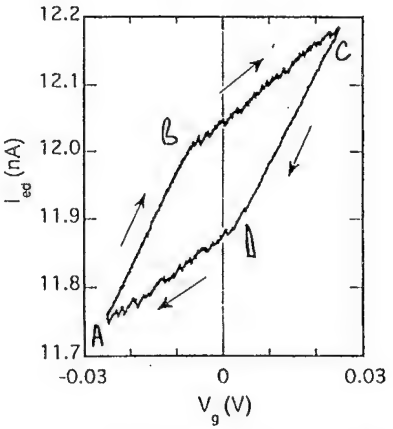


Figure 28. The electrometer current shows pronounced hysteresis as the gate voltage is cycled.

Eventually the voltage at the memory node will exceed the critical voltage of MTJ1 ($\sim e/2C_{MTJ}$) and current can flow through the SET allowing single electrons to charge the memory node. We have now reached point B and since the impedance of MTJ1 has been reduced a smaller fraction of the applied voltage appears at the memory node resulting in a pronounced change in gradient, which persists until we reach point C. If the gate voltage is now reduced the n electrons which have tunnelled onto the memory node are trapped there until the bias across MTJ1 is sufficient to overcome the Coulomb blockade in the reverse direction. This occurs at point D and the memory node can discharge by the time the gate voltage is reduced back to A. By cycling the gate voltage from D to A to B we can 'write' a Bit to the memory cell and by cycling from B to C to D we can remove it. After each half cycle we end up close to $V_g = 0$ V but the voltage at the memory node is different in each case.

The power dissipated by a single bit in a memory cell is proportional to the number of electrons which must be stored during each refresh cycle. For a state-of-the-art 16 MBit DRAM the capacitor of each element is ~ 35 fF (Rischetal, 1990) and will hold the order of 200,000 electrons when biased to 1 V and the total chip will dissipate heat of order 0.1 W. If the same technology is used to manufacture a 1 TBit (10^{12} Bits) and we simply scale the heat dissipation we expect a thermal output of more than 6kW which is comparable to a small electric heater. Nakazato *et al.* (1993) estimate that just 40 electrons are involved in the charging of their memory cell and if the technology can be developed to allow the ultra large scale integration of single electron transistors the reduction in heat dissipation will be a considerable advantage compared to conventional CMOS design.

REFERENCES

Averin, D. V., and Likharev, K. K., 1986, *J. Low Temp. Phys.* 62:345.

Averin, D. V., and Likharev, K. K., 1992, *in* "Single Charge Tunneling," ed. H. Grabert and M. H. Devoret (Plenum, New York).

Baranger, H. U., and Stone, A. D., 1989, *Phys. Rev. Letts.* 63:414.

Brown, E. R., *et al.*, 1989, *Appl. Phys. Letts.* 55:1777.

Datta, S., 1989, *Superlatt. and Microstucures* 6:83.

Esaki, L., 1958, *Phys. Rev.* 109:603.

Feng, Y., Thornton, T. J., Harris, J. J., and Williams, D., 1992, *Appl. Phys. Letts* 60:94.

Goodings, C. J., Cleaver, J. R. A., and Ahmed, H., 1992, *Electron. Letts.* 28:1535.

Grabert, H., 1991, *Z. Phys. B* 85:319.

Guéret, P., Blanc, N., Germann, R., and Rothuizen, H., 1992, *Phys. Rev. Letts.* 62:1896.

Ikoma, T., Odagiri, T., and Hirakawa, K., 1992, *in* "Quantum Effect Physics, Electronics and Applications" *IOP Conf. Ser.* 127, ed. by K. Ismail, T. Ikoma, and H. I. Smith (IOP Pub.).

Kouwenhoven, L. P., *et al.*, 1991, *Phys. Rev. Letts.* 67:1626.

Landauer, R., 1957, *IBM J. Res. Dev.* 1:223.

Likharev, K. K., 1987, *IEEE Trans. Mag.* 23:1142.

Likharev, K. K., 1990, *in* "Granular Nanoelectronics" *NATO ASI Series* 251:371, ed. D. K. Ferry, J. R. Barker, and C. Jacaboni (Plenum).

Luryi, S., 1985, *Appl. Phys. Letts.* 47:490.

Meirav, U., Kastner, M. A., and Wind, S. J., 1990 *Phys. Rev. Letts.* 65:771.

Mizuta, H., Goodings, C. J., Wagner, M., and Ho, S., 1992, *J. Phys. C: Condens. Matter* 4:8783.

Nakazato, K., Thornton, T. J., White, J., Ahmed, H., 1992 *Appl. Phys. Letts.* 61:3145.

Nakazato, K., Blaikie, R. J., Cleaver, J. R. A., and Ahmed, H., 1993, *El. Letts.* 29:384.

Odintsov, A. A., 1991, *Appl. Phys. Letts.* 58:2695.

Reed, M. A., *et al.*, 1988, *Phys. Rev. Letts.* 60:535.

Sakaki, H., 1980, *Jpn. J. Appl. Phys.* 19:L735.

Sakaki, H., 1989, *Jpn. J. Appl. Phys.* 28:L314.

Solls, F., Macucci, M., Ravaioli, U., and Hess, K., 1989, *Appl. Phys. Letts.* 54:350.

Spector, J., *et al.*, 1990a, *Appl. Phys. Letts.* 56:1290.
Spector, J., *et al.*, 1990b, *Appl. Phys. Letts.* 56:2433.
Tewordt, M., *et al.*, 1990, *J. Phys.:Condens. Matter* 2:8969.
Thornton, T. J., Pepper, M., Ahmed, H., Davies, G. J., and Andrews, D., 1986, *Phys. Rev. Letts.* 56:1198.
Van Wees, B. J., *et al.*, 1988, *Phys. Rev. Letts* 60:848.
Wharam, D. A., et al., 1988, *J. Phys. C: Solid State Phys.* 21:L209.
Wieck, A. D., and Ploog, K., 1990, *Surf. Sci.* 229:252.
Zimmerli, G., Kantz, R. L., and Martinis, J. M., 1992, *Appl. Phys. Letts.* 61:2616.

SOME RECENT DEVELOPMENTS IN QUANTUM TRANSPORT IN MESOSCOPIC STRUCTURES AND QUANTUM WELLS

L. Eaves, P. H. Beton, A. K. Geim and P. C. Main

Department of Physics
University of Nottingham
Nottingham NG7 2RD
U.K.

INTRODUCTION

This paper describes some recent developments in our understanding of quantum transport phenomena in mesoscopic semiconducting structures. There are already some excellent reviews on universal conductance fluctuations and quantum interference effects in mesoscopic metallic and semiconducting structures in the linear transport regime. Therefore this subject is dealt with only briefly in the following section, which also highlights some recent work on the magnetoresistance of semiconducting wires in the high magnetic field regime. The paper then goes on to describe the manifestation of mesoscopic effects in small-area resonant tunneling devices, which are inherently non-linear conductors, exhibiting strong deviations from Ohm's Law behavior. It is shown that in the voltage range below the threshold for conventional resonant tunneling into the quasi-bound continuum states of the quantum well, the current is entirely controlled by tunneling through states related to the presence of shallow donor impurities. These give rise to a current-voltage signature which varies from sample to sample. The number of such impurity states can be controlled by the introduction of intentional doping into the quantum well, combined with the appropriate design of the heterojunction layer and the choice of growth temperature. Resonant tunneling through deep impurities of this type can be used as a novel spectroscopic tool for probing the electronic states of a two-dimensional electron gas. At low temperatures the tunnel current shows evidence for singular behavior when the Fermi energy of the electrons in the emitter contact is resonant with an impurity. This Fermi edge singularity is a result of electron-electron interactions.

CONDUCTANCE FLUCTUATIONS IN THE LINEAR TRANSPORT REGIME

During the last decade, an enormous number of experiments and theoretical investigations have been performed on low dimensional and mesoscopic semiconducting structures. The terms "low dimensional" and "mesoscopic" are used rather loosely. "Low dimensional" refers to structures in which the electron motion is confined to less than three dimensions: e.g. the two dimensional electron gas (2DEG) in quantum well lasers or high electron mobility transistors; one-dimensional electron waveguides in ballistic point contacts and quantum wires; quantum dots in which electrons are quantum confined in all three dimensions. The term "mesoscopic" usually (although not uniquely) refers to structures which are considerably larger than the de Broglie wavelength of the carriers, but comparable in size to the quantum mechanical phase-breaking length, L_ϕ. In mesoscopic structures, which can be either truly metallic (e.g. Au wires) or semiconducting, one of the most important results has been the observation of a quantum contribution to the conductance, showing that the probability for an electron to be transmitted through the conductor depends not only on the number of scatterers, but also on the interference of partial electron waves propagating through the structure. The interference pattern and resulting transmission probability are sensitive to the positions of individual scatterers and displacement of even one of them can totally change the interference pattern and the conductance of the structure. Thus classically identical structures, with a slightly different microscopic distribution of impurities, have completely different quantum contributions to the conductance. The quantum contribution can be observed most easily by applying a weak magnetic field which changes the interference pattern, since an additional phase difference is introduced along the electron trajectories, due to the Aharonov-Bohm effect. In diffusive samples with many impurities, these changes in interference cause random fluctuations of the electrical conductance when an external parameter such as magnetic field or gate voltage is varied. In the case when the phase coherence persists over the entire volume of the structure, the fluctuations, sometimes termed universal conductance fluctuations (UCF), have magnitude DG of about e^2/h, which is the quantum of electrical conductance. For larger structures, the interference contributions from different phase-coherent units of size L_ϕ, are statistically averaged. As a result, conductance fluctuations average out in macroscopic samples and can be only observed in mesoscopic conductors having the size of a few L_ϕ. In semiconductors, L_ϕ is typically of the order of 1 mm at liquid helium temperatures. The term "universal" is applied since the amplitude and characteristic frequency of the fluctuations are characterized by the single parameter L_ϕ rather than the detailed form or composition of the conducting material.

UCF have been studied in great detail in the low magnetic field, dirty conductor limit, where $\omega_c\tau \ll 1$. In this case, the magnetic field only influences the quantum interference through the Aharonov-Bohm effect, but does not greatly affect the electron trajectories. This topic is discussed in an excellent review by Washburn and Webb (1992).

Timp *et al.* (1987, 1989) and others (Ford *et al.*, 1989; Taylor *et al.*, 1989; Ishibashi *et al.*, 1990; Bird *et al.*, 1991) investigated UCF at high magnetic fields. These investigations included measurements of electron transport in ballistic, sub-micron wires fabricated from a high-mobility two-dimensional electron gas (2DEG). In those structures the sample width w was much smaller than the electron mean free path l, and boundary scattering gave rise to a diffusive-like motion of electrons, which is necessary for the appearance of the UCF. At low magnetic fields, ballistic wires exhibit mesoscopic fluctuations, but these are usually suppressed at moderate magnetic fields and the quantum Hall effect is observed at still higher fields. This suppression of UCF arises from the suppression of diffusive motion when the electron-cyclotron orbit becomes smaller than the width of the wire; essentially, the electrons move in edge states with the suppression of backscattering.

Until recently, there has been little work on the regime $\omega_c\tau > 1$ in diffusive structures. This is not surprising, since the regime is usually inaccessible in traditional mesoscopic structures fabricated from "dirty" metallic films. However, heavily doped semiconductors can be employed for such experiments, and conductance fluctuations have been reported for $\omega_c\tau > 1$ in n^+ GaAs wires (Taylor *et al.*, 1989; Gallagher *et al.*, 1990) and low mobility 2DEGs in GaAs/AlGaAs heterostructures (Bykov *et al.*, 1990). The main problem in the study of UCF at high magnetic

fields is the rapid variation of the average magnetoresistivity due to the Shubnikov-de Haas effect. This Landau level effect is an unavoidable feature in the regime $\omega_c\tau > 1$, and normally conceals the UCF (Gallagher *et al.*, 1990; Bykov *et al.*, 1990).

In order to follow the true behavior of UCF in high magnetic fields, Geim *et al.* (1991, 1992) and Brown *et al.*, (1993) have employed a novel approach. They used highly diffusive semiconducting wires fabricated from a degenerate quasi-2DEG in n^+ GaAs. Here, quasi-2DEG refers to the fact that several (typically 4 or 5) subbands are occupied. This system represents an orthodox mesoscopic conductor with width $w \gg l$, but at the same time allows one to reach the $\omega_c\tau > 1$ regime at a reasonable low magnetic field, approximately 6 T. In addition, they employed a non-local geometry for the magnetoresistance measurements, which eliminated the effect of the Shubnikov-de Haas oscillations. They were thereby able to make a quantitative study of the UCF. The region of $\omega_c\tau > 1$ has also been investigated theoretically (Xiong and Stone, 1992) and it was suggested that the UCF amplitude quenches in high magnetic fields due to the very strong suppression of electron diffusion, with the consequent decrease of L_ϕ. This decrease of L_ϕ should manifest itself as a rapid increase of the characteristic period, ΔB, of the UCF in the magnetoresistance plots. This is indeed observed in experiments (Timp *et al.*, 1987, 1989; Geim *et al.*, 1991, 1992; Brown *et al.*, 1993) but there is considerable doubt whether the theory is a correct description of the results.

Typical experimental plots using the non-local geometry are shown schematically in Fig. 1 for an n^+ GaAs wire in both the local and non-local magnetoresistance configurations. It can be seen that whereas ΔB increases with increasing B, the amplitude of the UCF remains almost constant. This is in stark disagreement with the current theory of UCF in the high magnetic field regime, which predicts an exponential quenching of amplitude of the non-local UCF with increasing B ($\sim\exp(-L/L_\phi)$, where L is the length of the non-local probe). Brown *et al.* (1993) have been able to account for this qualitative discrepancy by considering the role of the edge of the sample in electron diffusion at high fields. At the edge of the sample, boundary diffusion occurs and this can dominate over the bulk contribution to quantum transport. Effectively, at high magnetic fields the sample consists of two phase-coherent parts: bulk and edge regions with differing values of L_ϕ. In high magnetic fields, the UCF from the small phase-coherent units in the bulk are rapidly averaged out and the fluctuations caused by the extended trajectories at the boundaries can dominate (back scattering at the boundaries is not totally suppressed, however, since boundary trajectories can be strongly coupled to the bulk and be eventually transmitted to the opposite edge; see Brown *et al.*, 1993). In this type of structure, truly "universal" scaling of the UCF amplitude and period breaks down, independent of either the geometry or type of measurement (Hall, local, non-local etc.) due to the presence of extended electron diffusion at the boundaries. Further theoretical work on this problem would be invaluable.

TUNNELING THROUGH ZERO DIMENSIONAL IMPURITY STATES IN MACROSCOPIC RESONANT TUNNELING DEVICES

Various authors have described experiments on the transport properties of laterally confined, sub-micron, resonant-tunneling diodes (RTD's) (Reed *et al.*, 1988; Guéret *et al.*, 1992; Tarucha *et al.*, 1990; Tewordt *et al.*, 1992; Su *et al.*, 1991; Dellow *et al.*, 1992). Invariably, the current-voltage, $I(V)$, characteristics of these devices displayed features additional to those seen in diodes with large cross sectional areas and the features have been attributed to either 0D quantisation (Reed *et al.*, 1988; Guéret *et al.*, 1992; Tarucha *et al.*, 1990; Tewordt *et al.*, 1992) or Coulomb blockade (Tewordt *et al.*, 1992; Su *et al.*, 1991). However, Dellow *et al.* (1992) have argued that similar features in $I(V)$ arise from tunneling through the bound states of individual donor impurities in the quantum well and the identification of the structure in $I(V)$ with the externally imposed lateral confinement is far from unambiguous (Beton *et al.*, 1992). In this section we show that it is possible to observe effects due to 0D states in resonant tunneling diodes (RTD's) which are not microscopic in lateral extent. Indeed, the existence of 0D states is a quite general property of any RTD containing impurities and the number of these states is related to the total number of impurities in the active region of the device. However, these states cannot be associated with single isolated donors.

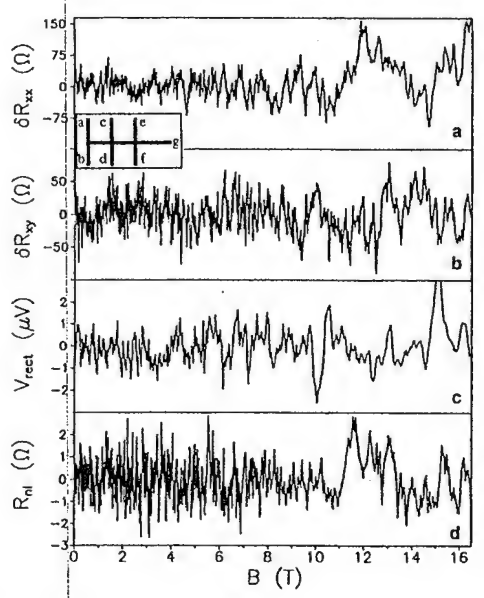


Figure 1. Universal conductance fluctuations at 4.2 K in different geometries for an n^+ GaAs wire with electron concentration $n = 1 \times 10^{24}$ m^{-3}, mobility $\mu = 0.18$ m^{2}V^{-1}s^{-1} and conducting thickness 30 nm with 4 electrically quantized 2D subbands occupied. The wire width is 350 nm. Adjacent pairs of probes (e.g. ab and cd) are separated by 1 mm and each contact probe has the same thickness and width as the main wire. We use the convention $R_{ijkl} = V_{kl}/I_{ij}$ so that in (a) $R_{xx} = R_{agce}$, in (b) $\delta R_{xy} = \delta R_{agcd} - B/ne$, and in (d) the non-local resistance $R_{nl} = R_{abcd}$. Panel (c) shows rectification fluctuations.

Figure 2 shows a schematic energy band diagram for a typical device under bias. Tunneling occurs from a two-dimensional electron gas (2DEG), formed in an accumulation layer near the left-hand emitter barrier, into the electron states in the well. A current flows when the energy of an electron in the 2DEG is resonant with a state in the quantum well (for a comprehensive review, see Chang *et al.*, 1991). The double barrier RTD's were grown by molecular beam epitaxy on n+GaAs substrates with substrate temperatures between 480 C and 550 C to inhibit donor segregation from the doped contacts into the active region of the device (Harris *et al.*, 1991). The center plane of the quantum well, width 9 nm, is doped with a narrow δ-layer (~1 nm) of Si donors with concentrations between $2 - 8 \times 10^{13}$ m^{-2} such that the mean separation of the donors is much greater than the Bohr radius in GaAs, 9.9 nm, and the impurity atoms can be regarded as isolated. The barrier thickness is 5.7 nm and there is a 10 nm - 20 nm spacer layer between each barrier and the more heavily doped contact regions. We also grew control samples in which the δ-layer of impurities had been omitted. Square mesas of side

lengths varying between 6 nm and 100 nm were fabricated using photolithography and dry or wet etching.

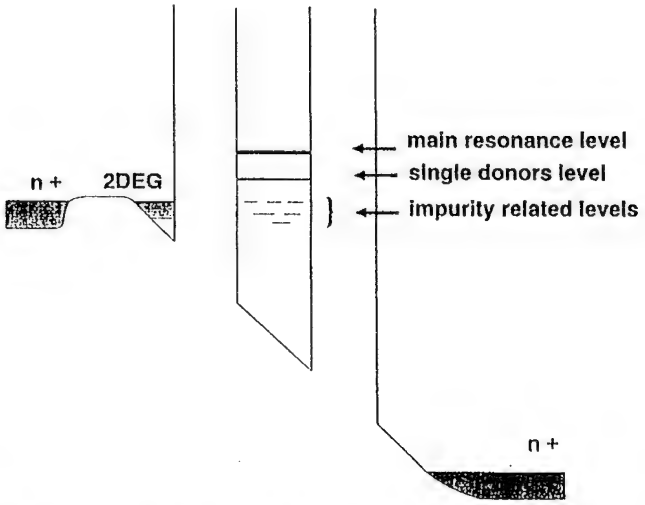


Figure 2. Energy band diagram of a typical resonant tunneling device. Tunneling occurs from a two-dimensional electron gas through the continuum (for the main resonance), single impurity levels or low energy, highly localized, impurity related levels.

It has already been established (Sakai *et al.*, 1993) in similar devices that the incorporation of donors into the quantum well of a large area RTD leads to a broad feature in $I(V)$ below the threshold for the main resonance. This feature is due to tunneling via the bound states of singly-charged isolated donors near the center of the well (see Fig. 2) which are lower in energy than the continuum, ie the main resonance level. For a 9 nm wide quantum well, these states have an energy around 12 meV below the continuum (Greene and Bajaj, 1983) and give rise to a feature at V = 90 mV. In Fig. 3 we show $I(V)$ at even lower bias voltages, near the onset of the tunnel current, at temperatures between 0.28 K and 10 K for a 6 nm device with 2×10^{13} m^{-2} donor concentration in the well. Forward bias is defined as a positive voltage applied to the top contact. The inset to Fig. 3 shows the main resonance in $I(V)$ for this device. In a control device with no donors in the well, the current is zero (< 0.1 pA) for biases up to ~90 mV. Clearly the presence of the donors leads to extra structure in I(V) at low biases. The extra structure occurs in both bias directions although the detail may differ slightly. Also, although the extra structure is qualitatively similar for all devices of the same size and doping, the detailed form is unique to a particular diode.

The presence of a tunneling current in a voltage range well below that for resonant tunneling into the continuum state of the well implies the existence of dopant-related tunneling channels, corresponding to localized states associated with the donors but with energies below that of the shallow donor level. The temperature dependence of the current near onset in Fig. 3 is thermally activated and can be written $I = I_0 f[\alpha e(V_t\text{-}V)/k_B T]$ where $f[x] = 1/(e^x + 1)$ is the Fermi function. I_0 and $\alpha = 0.27$ are constants and V_t is the threshold voltage, which is 31 mV in Fig. 3. The constant α is characteristic of the distribution of voltage across the device; it relates the applied voltage to the corresponding energy difference between the Fermi energy, E_F, in the emitter accumulation layer and the tunneling channel in the quantum well. The current onset is still activated at 70 mK and similar behavior has been seen at the current onset of most devices investigated provided the number of donor impurities is not too large. The sharpness of the onset implies that the tunneling is between the Fermi level, E_F, of the emitter accumulation layer and a state in the well which has an energy well below the continuum (ie the bound state of the quantum well) and a very narrow linewidth. The value of α allows us to convert the voltage across the device to the energy difference between E_F and the localized state. We find that the energy of the state is ~20 meV below the continuum, compared with ~12 meV for a single donor in a quantum well (Greene and Bajaj, 1983). Therefore, although we can be sure the states are donor-related, since their number is much reduced in an undoped sample, they are generally deeper than typical isolated, shallow hydrogenic donor levels as is shown schematically in Fig. 2.

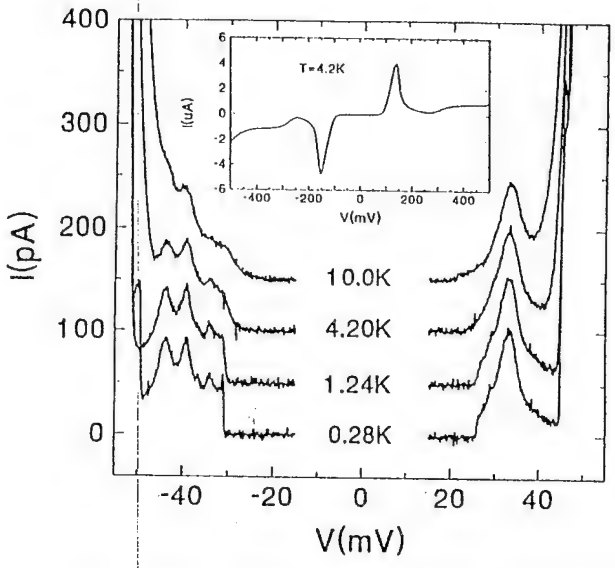


Figure 3. $I(V)$ characteristics at various temperatures at low bias for a device 6 nm across with 4×10^{13} m^{-2} donors incorporated at the center of the quantum well. Inset: $I(V)$ for the main resonance at T = 4.2 K.

The basic experimental results are the same for all devices with three general observations. First, the extra features in $I(V)$ are more extensive for larger device area at a given concentration of dopants. This is to be expected if we identify a feature in $I(V)$ with a particular localized state in the well. The larger the area, then the more such states we would expect. Secondly, in a device of a given lateral size the number of features increases in proportion to the number of donors incorporated in the well. Thirdly, the voltage range over which the sub-threshold features occur increases as the number of donors is increased. Note that this behavior is different to that of the peak due to tunneling through single donor states in these samples reported by Sakai *et al.* (1993). In that case the peak amplitude scales with the number of impurities but the voltage range is unaffected. These three observations are shown in Figs. 4 and

5. In Fig. 4, we show forward bias $I(V)$ characteristics at 0.3 K for three 12 mm devices with donor concentrations of (a) 0, (b) 4×10^{13} m^{-2} and (c) 8×10^{13} m^{-2} in the center of the well. All the qualitative features are similar to those of the 6 mm device but there is an obvious decrease in the number of sub-threshold features as the number of donors is reduced. For these devices the threshold for the continuum resonance is ~ 120 mV. Figure 5 shows $I(V)$ at 0.3 K for two devices of 100 mm lateral dimension, a size generally considered to be macroscopic, with donor concentrations (a) 0 and (b) 4×10^{13} m^{-2}. Again the effect of the donors is clear but for these large area diodes there are far more features in the undoped device than are visible in either of the two smaller devices as may be seen by comparing the (a) curves in Figures 4 and 5.

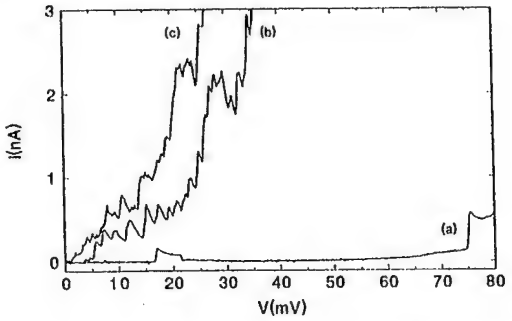


Figure 4. $I(V)$ characteristics at 0.3 K for three devices 12 mm across with donor concentrations of (a) 0, (b) 4×10^{13} m^{-2}, and (c) 8×10^{13} m^{-2}.

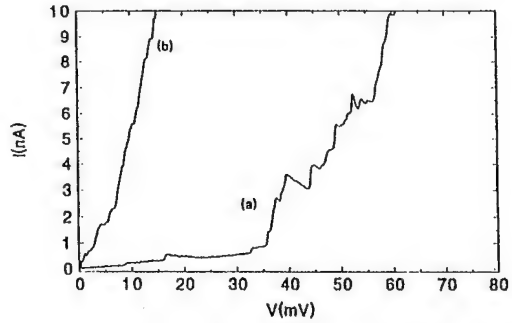


Figure 5. $I(V)$ characteristics at 0.3 K for two devices 100 mm across with donor concentrations of (a) 0 and (b) 4×10^{13} m^{-2}.

The characteristic shown in Fig. 4(a) is particularly interesting. Although the curve is for an undoped device there is a single feature which we estimate to be due to a localized state in the well with an energy approximately 25 meV below the continuum. This is a common but not universal feature of the control versions of the smaller area mesas. Most of these devices show no structure in $I(V)$ at all until very close to the threshold for the continuum resonance. However, a minority (~10-20 %) of them show single features which are similar to that shown in Fig. 4(c), both in terms of their shape and their temperature dependence. We have no firm identification of the origin of these single features: they are unlikely to be due simply to the presence of independent single donors which we would expect to be a maximum of ~12 meV below the continuum. Nevertheless we can assert that the large binding energy means that the states are localized on a length scale of a few nm and are discrete 0D states. Since we know that the onset of tunneling is still limited by k_BT even at $T = 70$ mK, the linewidth of the localized state is < 6 meV. The Heisenberg Uncertainty Principle, DE.t ~ h, relates the linewidth to the lifetime of the state; 6 meV is equivalent to ~ 1 ns which is consistent with the time for an electron to tunnel out of the well, estimated for the effective barrier width of our device. In addition, a tunneling time of 1 ns for a typical current of ~200 pA near onset indicates that the current is due to the passage of single electrons through the localized states.

Our results are a clear illustration of the observation of quantisation in all three dimensions in a transport experiment on macroscopic devices. The 0D states, far from being a unique property of sub-micron systems, we believe to occur in essentially all resonant tunneling devices of any size. In large area devices, particularly ones grown at relatively high temperatures, there are large numbers of these states and there may be a finite conductance, dI/dV, even at zero bias. However, there is no fundamental difference between devices of different area and our investigations cast some doubt on the earlier claims for the observation of lateral confinement and Coulomb blockade effects in sub-micron RTD (Reed *et al.*, 1988; Guéret *et al.*, 1992; Tarucha *et al.*, 1990; Tewordt *et al.*, 1992; Su *et al.*, 1991). Note also that the effect of impurities is far more evident in tunneling phenomena than in lateral transport through quantum dots formed in 2DEG's (see, e.g., Geeligs *et al.*, 1993). In the latter case, the relatively high electron concentration screens the impurities and the potential fluctuations are much smaller. In contrast, in a tunneling device at voltages below the threshold for the main resonance, the electronic charge density is very low both in the barriers and in the well as a result of the evanescent character of the electron wavefunction.

RESONANT TUNNELING SPECTROSCOPY OF A FERMI EDGE SINGULARITY

In Fig. 6(a) we show in more detail the isolated peak of Fig. 4(a) and Fig. 6(b) is the current onset for a δ-doped RTD with 2×10^{13} m^{-2} Si donors in the quantum well. The unexpected feature in the observed $I(V)$ dependences is the singular enhancement of tunneling near the threshold, when the localized state is resonant with the emitter Fermi energy. The characteristic width of the peaks at threshold biases can be as small as 0.2 mV at the lowest temperature (e.g. see the marked feature in Fig. 6(b)). The low voltage edge of each step is thermally activated down to 70 mK indicating that the 2DEG remains in thermal equilibrium with the main heat bath. In general, as in Fig. 6(b), there is some additional oscillatory structure within the step at voltages above the threshold voltage V_{th}. However, in contrast to the singularity this structure does not depend on temperature. A Fermi-edge singularity of this type is seen in all devices at temperatures below 1 K and we attribute it to the Coulomb interaction between the tunneling electron on the localized site and the Fermi sea of the 2DEG (there is a vast literature on this topic; some recent papers are Skolnick *et al.*, 1987; Lee *et al.*, 1987; Livescu *et al.*, 1988; Chen *et al.*, 1992; Hawrylak, 1992). The localized state occurs at an energy significantly below that expected for an isolated donor in the quantum well and is attributed to a closely-spaced donor pair.

The general behavior of the tunnel current, i.e. without the singularity, can be understood as follows. Under a typical applied bias of tens of mV, the collector barrier is lower than the emitter barrier. Also the 3D density of states in the collector available to the electron tunneling from the localized state is larger than the equivalent density of states in the emitter 2DEG. Consequently, the current is limited by tunneling through the emitter barrier and the states in the quantum well are empty most of the time. We estimate the escape rate τ_e^{-1} from the impurity into

the collector contact to be considerably larger than the tunneling rate τ_e^{-1} from the 2DEG into the impurity-related site. The latter rate can be found from a value of the single electron current through an impurity, $I \cong 100$ pA, yielding $\tau_e = e/I \cong 2$ ns. This is consistent with the barrier height and thickness. As the bias increases, the impurity level moves downwards relative to the energy of the emitter 2DEG (see Fig. 2) and the tunnel current exhibits a step increase when the localized state coincides with the highest filled state, i.e. the Fermi level of the 2DEG. As the voltage is increased further and the energy of the 0D state becomes lower than the lowest energy state of the emitter, no states are available for resonant tunneling and the current falls sharply (see Fig. 6(a)). However, this sharp fall-off is not common to all devices. In Fig. 6(b), a second impurity channel comes into resonance with the 2DEG (at $V \cong 50$ mV) before the first channel has passed away.

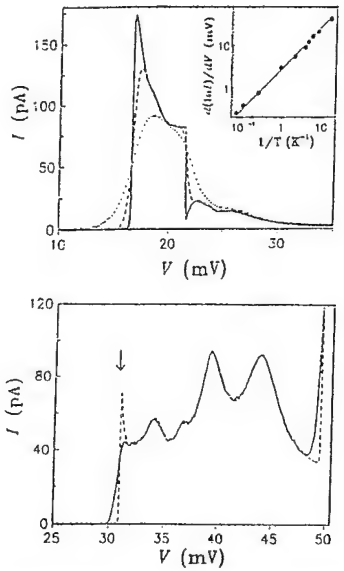


Figure 6. Detailed $I(V)$ characteristics at low biases when the first localized level is resonant with the emitter 2DEG. (a) (top) - the same device as in Fig. 2 at three different temperatures of 70 mK (solid line), 1.3 K (dashed line) and 5 K (dots). Inset - temperature dependence of the logarithmic slope of the tunnel current near the threshold voltage. (b) (below) - another device 6 mm across at 70 mK (dashed line) and 1.2 K (solid line).

For quantitative analysis, we discuss first the behavior expected for impurity-assisted tunneling of non-interacting electrons. The tunnel current is given by

$$I = \frac{e}{\hbar}\Gamma(\varepsilon_i)\theta(\varepsilon_i) f(\varepsilon_i) \ ,$$

where $f(\varepsilon)=\{1+\exp[(\varepsilon-\varepsilon_F)/k_BT]\}^{-1}$ is the Fermi-distribution function, $\theta(\varepsilon)$ is the unit step-function and ε_i is the energy of the impurity state measured from the bottom of the 2DEG subband. For the case of tunneling from a 2DEG, the tunneling coefficient Γ can be written as

$$f(\varepsilon_i) = t\exp(-\varepsilon_i/\varepsilon_o) \ , \qquad (2)$$

where ε_o is the binding energy of the localized state and t is a coefficient which includes parameters of the localized state and the tunnel barrier but is independent of the kinetic energy ε_i of tunneling electrons within a 2DEG subband. Near the onset of the tunnel current, (1) varies as the Fermi function which fits very well the observed $I(V)$ characteristics and their temperature dependence. This allows us to convert the voltage across the device into the energy difference between the impurity state and ε_F. We use $\varepsilon_F-\varepsilon_i = \alpha e(V-V_{th})$ where the constant α is characteristic of the distribution of electrostatic potential across the device. Experimental curves yield $\alpha = 0.25\pm0.05$ for all devices. The inset in Fig. 6(a) shows an example of temperature dependence of the tunnel current below the threshold at biases when $I \propto f(\varepsilon_i) \cong \exp[(\varepsilon_F-\varepsilon_i)/k_BT)]$ and one can write $d\ln(I)/dV \cong \alpha/k_BT$. The linear dependence of the logarithmic slope down to the lowest temperature of 70 mK indicates that the localized state has a very narrow linewidth. We note that the sharp fall of the current which is seen at $V \cong 22$mV in Fig. 6(a) is not a common feature for the isolated peaks and, usually, the decrease is much smoother.

Equation (2) shows that, for non-interacting electrons, the tunnel current within the step varies only on the energy scale of the binding energy $\varepsilon_o \geq 13$ meV. On the other hand, typical values of the Fermi energy in the emitter accumulation layer for the first few steps in the I-V characteristics are between 1-4 meV in our experiment, corresponding to 4-20 mV in bias. Therefore, according to (1) and (2), variation of the current within the step has to be small. the dashed line in Fig. 7 shows the $I(V)$ characteristic for non-interacting electrons, calculated using (1) and (2) for the sample in Fig. 6(a). The binding and Fermi energies were estimated using the value of $a \cong 0.2$ for this sample, found by fitting the temperature dependence of the current onset. From Fig. 7 we conclude that the observed singularity in the tunnel current cannot be explained within a model involving only non-interacting electrons. Therefore, we attribute the FES to the influence of the electron-electron interaction.

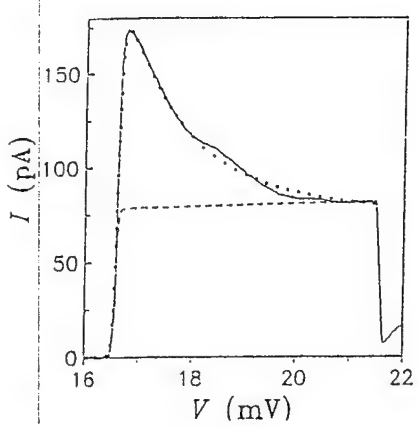


Figure 7. Comparison of the observed singularity with theory. The solid line is the experimental curve at 70 mK for the same device as in Fig. 6(a). The dotted curve is the behavior expected from the theory of Matveev and Larkin (1992). The dashed curve: if the electron-electron interaction is neglected, the tunnel current within the step exhibits only a very small increase with increasing bias.

Three models leading to a FES for the case of impurity assisted tunneling have been considered recently. First, the interaction between conduction electrons in the emitter 2DEG yields a logarithmic singularity in the tunneling density of states (Altshuler and Aronov, 1985). Secondly, repulsion between electrons with the opposite spins on the impurity site may give rise to a Kondo resonance (Ng and Lee, 1988; Glazman and Raikh, 1988; Glazman and Matveev, 1988; Hershfield *et al.*, 1991; Yeyati *et al.*, 1993; Ralph and Burman, 1992). Finally, the interaction between an electron on the impurity site and the Fermi sea in the emitter contact may cause the MND singularity (Matveev and Larkin, 1992). The first effect is important if the electron mean free path is short, $l \sim \lambda_F$, but is expected to be negligibly small for our 2DEG at the emitter interface. In addition, we would expect a negative contribution to the tunnel current near ε_F rather than the increase which is observed. The Kondo resonance also leads to $I(V)$ qualitatively different from that observed [for the corresponding $I(V)$ characteristics, see Bird *et al.*, 1991]. The FES in our experiments is very similar to the behavior predicted by Matveev and Larkin (1992) (ML). The singularity originates from extra tunneling processes due to the Coulomb interaction between the fluctuating charge on the localized site and the Fermi sea in the contacts. The interaction allows an electron to violate the requirement of energy conservation between its initial and final states in the tunnel process. In addition to the direct tunneling, the electron can tunnel into the localized site from an initial state which does not participate in tunneling in a model of non-interacting particles. The difference in the energies is transferred to or from the Fermi sea. A singularity arises at ε_F because scattering processes with small energy transfer are most effective (Fermi's golden rule) while the Pauli principle allows them only near the Fermi energy. The ML theory is basically analogous to the Mahan problem which considers the interaction between the Fermi sea and a positively charged virtual hole. The calculations take into account all many-body processes such as shake-up of the Fermi sea, excitonic effects, etc.

The ML theory yields a power-law singularity of the form

$$I \propto (\varepsilon_f - \varepsilon_i)^{-\beta}\, \theta(\varepsilon_f - \varepsilon_i) \ , \qquad (3)$$

$$\beta \cong 3\pi (k_F d)^{-1}/4 \ , \qquad (4)$$

where $k_F = 2\pi/\lambda_F$ is the Fermi wavevector and d is the distance from the plane of the 2DEG to the localized site. The parameter β is characteristic of the strength of the Coulomb interaction and can be found directly from the experimental data. We estimate $d \cong 25$ nm assuming the Fang-Howard approximation for the emitter 2DEG and that the localized states are in the middle of the quantum well. For the first few steps which occur at biases V between 15-90 mV, we found β between 0.1 and 0.3.

The interaction lasts a finite time τ_c, before an electron escapes from the impurity state into the collector contact, and this leads to smearing of the singularity on the energy scale $\Gamma_c = \hbar/\tau_c$. The smearing is described by adding the imaginary part $i\Gamma_c$ to ε_i in (3), leading to the expression (Matveev and Larkin, 1992)

$$I \propto \left(\sqrt{(\varepsilon_F - \varepsilon_i)^2 + \Gamma_c^2}\right)^{-\beta} \left(\frac{\pi}{2} + \arctan\frac{\varepsilon_F - \varepsilon_i}{\Gamma_c}\right) \theta(\varepsilon_F - \varepsilon_i) \ .$$

Referring to the estimates for the ratio between the tunneling rates τ_c^{-1} and τ_c^{-1} in our structures, we expect Γ_c to be of the order 0.1 meV ($\tau_c \cong 10$ps).

To describe the observed form of FES we assume that the net current includes both single-particle and many-body contributions given by (1) and (5), respectively. The absolute value of the many-body current in (3-5) is unknown and used as a fitting parameter. Also, we vary Γ_c around 0.1 meV to obtain the best agreement with the experimental data. Figure 7 shows the best fit to the low-temperature $I(V)$ characteristic from Fig. 6(a). The coefficient β is $\cong 0.22$ for this sample and the fit yields $\Gamma_c \cong 0.2$ meV. For other samples, the singularities are also described by values of Γ_c close to 0.1 meV (within a factor of two).

For completeness, to describe the temperature smearing at the onset of tunneling in Fig. 7, we have multiplied (5) by the Fermi function $f(\varepsilon_i)$, instead of using the theta-function as in Matveev and Larkin (1992). For higher temperatures, when $k_B T \cong \Gamma_c$, the smearing of FES is due to the effect of both temperature and Γ_c. Essentially, the singularity increases in size with decreasing temperature until it is limited by Γ_c. Further reduction in the temperature results in only minor modifications to the peak, although the current onset is still thermally activated. Although Fig. 7 shows quantitative agreement between the experiment and ML theory, we note that (3-5) are derived for biases close to the threshold. In addition, the numerical coefficient $3\pi/4$ in (4) is valid for the case $k_F d \gg 1$, while we deal with the situation when the interaction is strong and $k_F d \cong 1$. Further details of the experiment can be found in Geim *et al.*, (1994).

CONCLUSION

We have briefly reviewed some recent developments in the study of conductance fluctuations in the magnetoresistance of doped semiconducting wires of sub-micron dimensions at high magnetic fields. A related effect in the current-voltage characteristics of small area resonant tunneling diodes is examined and is shown to be due to tunneling through discrete levels associated with donor impurities. This impurity-related tunneling problem can be used to examine, for the first time, a Fermi edge singularity effect in quantum transport.

ACKNOWLEDGEMENTS

This work is supported by SERC/EPSRC (U.K.). PHB and LE wish to thank the Royal Society (U.K.) for financial support. We are grateful to Drs. G. Hill and M. A. Pate for device lithography and to Dr. M. Henini for growing the layers.

REFERENCES

Altshuler, B. L., and Aronov, A. G., 1985, *in* "Electron-Electron Interactions in Disordered Systems," Ed. by A. L. Efros and M. Pollak, Elsevier Science:Amsterdam.

Beton, P. H., Eaves, L., and Main, P. C., 1992, *Phys. Rev. Lett.* 69:2995.

Bird, J. P., *et al.*, 1991, *J. Phys. Cond. Matter* 3:2897.

Brown, C. V., Geim, A. K., Foster, T. J., Langerak, C. J. G. M., and Main, P. C., 1993, *Phys. Rev. B* 47:10935.

Brykov, A. A., *et al.*, 1990, *Superlatt. Microstruc.* 10:287.

Chang, L. L., Mendez, E. E., and Tejedor, C., Eds., 1991, "Resonant Tunneling in Semiconductors-Physics and Applications," Plenum:New York, NATO ASI Series B, vol. 277.

Chen, W., *et al.*, 1992, *Superlatt. Microstruc.* 4:4237.

Dellow, M. W., Beton, P. H., Langerak, C. J. G. M., Foster, T. J., Main, P. C., Eaves, L., Henini, M., Beaumont, S. P., and Wilkinson, C. D. W., 1992, *Phys. Rev. Lett.* 68:1754,

Ford, C. J. B., *et al.*, 1989, *Appl. Phys. Lett.* 54:21.

Gallagher, B. L., *et al.*, 1990, *Phys. Rev. Lett.* 64:2058.

Geerligs, L. J., Harmans, C. J. P. M., and Kouwenhouven, L. P., Eds., 1993, "The Physics of Few Electron Nanostructures," North-Holland:Amsterdam, *Physica B*, vol. 189.

Geim, A. K., Main, P. C., Beton, P. H., Streda, P., Eaves, L., Wilkinson, C. D. W., and Beaumont, S. P., 1991, *Phys. Rev. Lett.* 67:3014.

Geim, A. K., Main, P. C., Beton, P. H., Streda, P., Eaves, L., Wilkinson, C. D. W., and Beaumont, S. P., 1992, *Phys. Rev. Lett.* 69:1248.
Geim, A. K., Main, P. C., La Scala, N., Jr., Eaves, L., Foster, T. J., Beton, P. H., Sakai, J. W., Sheard, F. W., and Henini, M., 1994, *Phys. Rev. Lett.* 72:2061.
Glazman, L. I., and Matveev, K. A., 1988, *Pis'ma Zh. Eksp. Teor. Fiz.* 48:403 [transl. *in JETP Lett.* 48:445].
Glazman, L. I., and Raikh, M. E., 1988, *Pis'ma Zh. Eksp. Teor. Fiz.* 47:378 [transl. in *JETP Lett.* 47:452].
Greene, R. L., and Bajaj, K. K., 1983, *Sol. State Commun.* 45:825.
Guéret, P., Blanc, N., Germann, R., and Rothuisen, H., 1992, *Phys. Rev. Lett.* 68:1986.
Harris, J. J., Clegg, J. B., Beall, R. B., Castagne, J., Woodbridge, K., and Roberts, C., 1991, *J. Cryst. Growth* 111:239.
Hawrylak, P., 1992, *Phys. Rev. B* 45:8464.
Hawrylak, P., 1993, *in* "Optical Phenomena in Semiconductor Structures of Reduced Dimensions," Ed. by D. J. Lockwood and A. Pinczuk, NATO ASI Series E, 248:295.
Hershfield, S., Davies, J. H., and Wilkins, J. W., 1991, *Phys. Rev. Lett.* 67:3720.
Ishibashi, K., *et al.*, 1990, *Surf. Sci.* 228:286.
Lee, J. S., *et al.*, 1987, *Semicon. Sci. Technol.* 2:675.
Livescu, G., *et al.*, 1988, *Superlatt. Microstruc.* 4:359.
Matveev, K. A., and Larkin, A. I., 1992, *Phys. Rev. B* 46:15337.
Ng, T. K., and Lee, P. A., 1988, *Phys. Rev. Lett.* 61:1768.
Ralph, D. C., and Buhrman, R. A., 1992, *Phys. Rev. Lett.* 69:2118.
Reed, M. A., Randall, J. N., Aggarwal, R. J., Matyi, R. J., Moore, T. M., and Wetsel, A. E., 1988, *Phys. Rev. Lett.* 60:535.
Sakai, J. W., Fromhold, T. M., Beton, P. H., Eaves, L., Henini, M., Main, P. C., and Hill, G., 1993, *Phys. Rev. B* 48:5664.
Skolnick, M. S., *et al.*, 1987, *Phys. Rev. Lett.* 58:2130.
Su, B., Goldman, V. J., Santos, M., and Cunningham, J. E., 1991, *Appl. Phys. Lett.* 58:747.
Tarucha, S., Hirayama, Y., Saku, T., and Kimura, T., 1990, *Phys. Rev. B* 41:5459.
Taylor, R. P., *et al.*, 1989, *J. Phys. Cond. Matter* 1:10413.
Tewordt, M., Martin-Moreno, L., Nicholls, J. T., Pepper, M., Kelly, M. J., Law, V. J., Ritchie, D. A., Frost, J. E. F., and Jones, G. A. C., 1992, *Phys. Rev. B* 45:14407.
Timp, G., *et al.*, 1987, *Phys. Rev. Lett.* 59:732.
Timp, G., *et al.*, 1989, *Phys. Rev. B* 39:6227.
Washburn, S., and Webb, R. A., 1992, *Rep. Prog. Phys.* 55:1311.
Xiong, S., and Stone, A. D., 1992, *Phys. Rev. Lett.* 67:3014.
Yeyati, A. L., Martin-Rodero, A., and Flores, F., 1993, *Phys. Rev. Lett.* 71:2991.

INTERACTING AND COHERENT TIME-DEPENDENT TRANSPORT IN SEMICONDUCTOR HETEROSTRUCTURES

Antti-Pekka Jauho

Mikroelektronik Centret
Denmark's Technical University
DK-2800 Lyngby, Denmark

INTRODUCTION

The hallmark of mesoscopic phenomena is the phase coherence of the charge carriers, which is maintained over a significant part of the transport process. The interference effects resulting from this phase coherence are reflected in a number of experimentally measurable properties. For example, phase coherence is central to the Aharonov-Bohm effect, Universal conductance fluctuations, (for a series of review articles, see Altshuler *et al.*, 1991) and weak localization (reviewed by, e.g, Lee and Ramakrishnan, 1985), and can be affected by external controls such as temperature or magnetic field. The study of stationary mesoscopic physics is now a mature field, and in these notes we describe recent developments in a rapidly growing research area, where one uses an alternative way of affecting the phase coherence: external *time-dependent* perturbations. The interplay of external time-dependence and phase coherence can be phenomenologically understood as follows. If the single-particle energies acquire a time dependence, then the wave functions have an extra phase factor, $\psi \propto \exp\left[-i\int^{t} dt'\varepsilon(t')\right]$. For a uniform system such an overall phase factor is of no consequence. However, if the external time dependence is different in different parts of the system, and the particles can move between these regions (without being 'dephased' by inelastic collisions), the phase difference becomes important.

The interest in time-dependent mesoscopic phenomena stems from recent progress in several experimental techniques (see, for example, Kirk and Reed, 1992). Time dependence is a central ingredient in many different experiments, of which we mention the following: (i) *Single-electron pumps and turnstiles.* Here time-modified gate signals move electrons one by one through a quantum dot, leading to a current which is proportional to the frequency of the external signal. These structures have considerable importance as current standards. The coulombic repulsion of the carriers in the central region is crucial to the operational principle of these devices, and underlines the fact that extra care must be paid to interactions when considering time-dependent transport in mesocopic systems. (ii) *ac response and transients in resonant-tunneling devices.* Resonant-tunneling devices (RTD) have a number of applications as high-frequency amplifiers or detectors. For the device engineer a natural approach would be to model these circuit elements with resistors, capacitances, and inductors. The question then arises as to what, if any, are the appropriate 'quantum' capacitances and inductances one should ascribe to these devices. Answering this question requires the use of time-dependent quantum-transport theory. Similar requirements are valid, if one wishes to study the *noise* properies of tunneling devices. (iii) *Interaction with laser fields.* Ultrashort laser pulses allow the study of short-time dynamics of charge carriers. Further, the extremely strong ac fields which may be generated with free-electron lasers, provide a new and fascinating way of affecting the motion of charge carriers in semiconductor heterostructures. Here again, coherence and time dependence combine with the necessity of treating interactions.

A rigorous discussion of transport in an interacting mesoscopic system requires a formalism which is capable of including explicitly the interactions. Obvious candidates for such a theoretical tool are various techniques based on Green's functions. Since many problems of interest involve systems far from equilibrium, we cannot use linear-response methods, such as those based on the Kubo formula, but must use an approach capable of addressing the full nonequilibrium situation. The nonequilibrium Green's function techniques, as intitiated by Schwinger and Martin (1959,1961), and developed further by Kadanoff and Baym (1962), and by Keldysh (1964), have during the recent years gained increasing attention in the analysis of transport phenomena in mesoscopic semiconductor systems. Many of the recent papers follow the line of thinking suggested by Caroli and co-workers (1971a, 1971b, 1972). In particular, the steady-state situation has been adressed by a large number of papers (a few representative papers are included in the following list: Meir and Wingreen, 1992; Davies *et al.*, 1993; Anda and Flores, 1991; Hershfield *et al.*, 1991; Meir *et al.*, 1993; Lake and Datta, 1992). Among the central results obtained in these papers is that under certain conditions (to be discussed below) a Landauer (1957, 1970) type conductance formula can be derived. This is quite appealing in view of the wide-spread succes of conductance formulas in the analysis of transport in mesoscopic systems.

Considerably fewer studies have been reported where an explicit time dependence is an essential feature. We are aware of an early paper in surface physics (Blandin *et al.*, 1976), but only in the recent past have groups working in mesoscopic physics addressed this problem (see, for example, Chen and Ting, 1991; Langreth and Nordlander, 1991; Wingreen *et al.*, 1993; Bruder and Schoeller, 1994; Runge and Ehrenreich, 1992a, 1992b; Fu and Dudley, 1993; Pastawski, 1992). The first part of these notes is a brief review of the techniques and results of these papers, concentrating mainly on our own work (Wingreen *et al.*, 1993; Jauho *et al.*, 1994). The main formal result of these papers is a general expression for the time-dependent current flowing from non-interacting leads to an interacting region. The time-dependence enters through the self-consistent parameters defining the model. Under certain restrictions, to be specified below, we obtain a Landauer-like formula for the *time-averaged* current. For illustrative purposes we give a number of results for an exactly solvable non-interacting resonant-tunneling model, which displays an interesting, and experimentally measurable, nonadiabatic behavior. We also establish a link to a number of other recent results, obtained with other techniques.

The second part in these lectures focuses on Bloch oscillations, Wannier-Stark Ladders, and Zener tunneling. High-quality samples, based on semiconductor heterostructures, have resulted in an explosion of research activity in this area. Ever since the early studies by Bloch (1928) and Zener (1934), according to which a free electron in a periodic crystal, moving under the influence of a steady electric field E, should execute oscillations with frequency $\omega_B = eEd/\eta$ (here d is the lattice periodicity), speculations have been made of a high-frequency oscillator based on these simple ideas. In normal bulk material phase-braking collisions occur so frequently, that coherent motion is not possible

for a full Bloch period $T_B = 2\pi/\omega_B$, and consequently Bloch oscillations have evaded experimental observation until recently. The situation is different, however, in high-quality semiconductor superlattices, as pointed out by Esaki and Tsu (1970). Here the increased lattice periodicity leads to less stringent condtions for the observability of coherent charge oscillations, and consequently the last few years have witnessed the first observations of these oscillations (Leo *et al.*, 1992; Waschke *et al.*, 1993). While the experimental results are new, and perhaps their interpretation is not entirely straightforward, it is nevertheless clear that refined techniques may lead to many more interesting observations. An indication of this trend is the recent experiment by Guimaraes *et al.* (1993), where a strong ac field, originating from a free-electron laser, when focused on a superlattice, gives indications of photon-assisted resonant tunneling. This is a precursor for the experimental verification of another long-standing theoretical prediction: possibility to build lasers based on resonant transfer between charge carriers in coupled quantum wells (Kazarinov and Suris, 1972). Instead of applying the nonequilibrium Green's function methods described above, we address this problem complex by studying the equations-of-motion of the density matrix. This method can occasionally lead more directly to manageable expressions than the full nonequilibrium Green's function techniques. In particular, if scattering is entirely neglected (or described by a simple phenomenological model), as is done below, the equations-of-motion for the multi-band density matrix can be directly integrated. The knowledge of $\rho_{\alpha}(k,t)$ gives us the possiblity to discuss Zener tunneling and band collapse (Holthaus, 1992a, 1992b; Holthaus and Hone, 1993) in *time-domain*, which may lead to significant new insights.

As mentioned above, these notes are divided in two parts, one with focuses on Green's function methods, and another which applies the density matrix technique. Each part can be studied independently of the other, and both parts include a short introductory part, which reviews the basic physics and introduces the main technical results required as a prerequisite.

NONEQUILIBRIUM GREEN'S FUNCTIONS APPLIED TO MESOSCOPIC TRANSPORT

Theoretical background

The most important result of the formal theory of nonequilibrium Green's functions is that the perturbation expansion has precisely the same structure as the $T = 0$ equilibrium expansion (modern reviews of the theory can be found in, e.g. Langreth, 1976; Rammer and Smith, 1986; Jauho, 1989, 1991; Haug and Jauho, 1994). Instead of a time-ordered function, one works with the contour-ordered Green's function,

$$G(\tau,\tau') = -i\left\langle T_C\{\psi(\tau)\psi^+(\tau')\}\right\rangle , \tag{1}$$

where the complex time-variables τ,τ' live on the complex contour C shown in Fig. 1. The contour ordering operator T_C orders the operators following it in the contour sense: operators with time labels later on the contour are moved left of operators of earlier time labels. The physical reason behind the double-time formulation is that in nonequilibrium theory one cannot assume that the system returns to its ground state (or a thermodynamic equilibrium state at finite temperatures) as $t \to +\infty$. Irreversible effects break the symmetry between $t = -\infty$ and $t = +\infty$, and this symmetry is heavily exploited in the derivation of the equilibrium perturbation expansion. In nonequilibrium situations one can circumvent this problem by allowing the system to evolve from $-\infty$ to the moment of interest (for definiteness, let us call this instant t_0), and then continues the time evolvment from $t = t_0$ back to $t = -\infty$ (Schwinger (1959)). When dealing with quantities which depend on two time variables, such as Green's functions, the time evolution must be continued to the later time. The advantage of this procedure is that all expectation values are defined with respect to a well defined state, i.e. the state in which the system was prepared in the remote past.

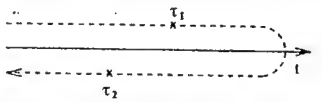


Figure 1. The complex-time contour on which nonequilbrium Green's function theory is constructed. In the contour sense, the time τ_1 is earlier than τ_2 even though its real time projection appears larger.

Summarizing, once the self-energy functional $\Sigma = \Sigma[G]$ has been specified, the contour-ordered Green's function obeys formally the same Dyson equation as in T=0 theory,

$$G = G_0 + G_0 \Sigma G , \tag{2}$$

with the modification that internal time-integrations run along the complex contour shown in Fig.1. The contour-ordered Green's function contains information of two kinds of Green's functions: correlation functions, or 'lesser' functions,

$$G^<(1,1') = i\left\langle \psi^+(1')\psi(1)\right\rangle , \tag{3}$$

and retarded/advanced functions:

$$G^r(1,1') = -i\theta\left(t_1 - t_{1'}\right)\left\langle\left\{\psi(1),\psi^+(1')\right\}\right\rangle , \tag{4a}$$

$$G^a(1,1') = i\theta\left(t_{1'} - t_1\right)\left\langle\left\{\psi(1),\psi^+(1')\right\}\right\rangle . \tag{4b}$$

Here we employ a shorthand notation for the arguments of the fermion field operators: $1 \equiv \left(\mathbf{x}_1, t_1\right)$. The curly brackets indicate an anticommutator, and for bosons one would use a commutator. Roughly speaking, the lesser function contains information about the distribution of the particles (and quantities derivable from the distribution, such as density and current), while the retarded/advanced functions contain information about densities of states, life-times and scattering rates. In order to find equations-of-motion for the nonequilibrium Green's functions, one needs a method to extract the retarded and lesser functions from the complex-time contour Dyson equation (2), which may involve several internal time-integrations. A particulary convenient method is due to Langreth (1976), which can be summarized as follows. If one has an expression $A = \int BC$ on the contour (this is the generic type of term encountered in the perturbation expansion), then the retarded and lesser components are given by

$$A^r(t,t') = \int dt_1 B^r\left(t,t_1\right)C^r\left(t_1,t'\right) , \tag{5}$$

$$A^<(t,t') = \int dt_1\left[B^r\left(t,t_1\right)C^<\left(t_1,t\right) + B^<\left(t,t_1\right)C^a\left(t_1,t\right)\right] . \tag{6}$$

These results are readily generalized to products involving three (or more) Green's functions or self-energies. Applying these rules, one can derive from the contour-ordered Dyson equation either the Keldysh equation,

$$G^< = (1+G^r\Sigma^r)G_0^<(1+\Sigma^a G^a)+G^r\Sigma^< G^a\ , \tag{7}$$

or the (generalized) Kadanoff-Baym equation,

$$[G_0^{-1}-U,G^<]-[\Sigma,G^<]-[\Sigma^<,G]=\tfrac{1}{2}\{\Sigma^<,G^>\}-\tfrac{1}{2}\{\Sigma^>,G^<\}. \tag{8}$$

The square brackets indicate a commutator, and we have defined

$$\Sigma=\tfrac{1}{2}(\Sigma^r+\Sigma^a),\quad G=\tfrac{1}{2}(G^r+G^a), \tag{9}$$

and the symbol U contains the driving forces, heterojunction conduction-band-edge potentials, and other one-body potentials, e.g. self-consistent Hartree terms. To obtain a closed set of equations the Keldysh equation, or the Kadanoff-Baym equation must be supplemented with the Dyson equations for the retarded and advanced Green's functions:

$$G^{r,a}=G_0^{r,a}+G_0^{r,a}UG^{r,a}+G_0^{r,a}\Sigma^{r,a}G^{r,a}. \tag{10}$$

The equations (7-10) are exact (given that a self-energy can be defined), and form the starting point for the theory.

Many different applications exist in the literature, and it is beyond the scope of these lectures to give an extensive review. The following short list should be viewed as a starting point for a more thorough study of the literature. The linear response (in driving fields) studied with the aid of the quantum kinetic equations is on a firm basis, and has been reviewed by Mahan (1984,1987). Equivalence of the quantum kinetic equations and Kubo formula (within linear response) has been demonstrated (Chen and Su, 1989). An interesting application is the study of disordered systems within the formalism: Hershfield and Ambegaokar (1986), Strinati *et al.* (1989), and Suhrke and Wilke (1992) have derived quantum kinetic equations to treat localization effects. Effects that are *nonlinear* in the driving fields have for a long time been a central activity (Jauho and Wilkins, 1982, 1984; Vasko, 1983; Lipavsky *et al.*, 1986; Khan *et al.*, 1987; Reggiani *et al.*, 1987). In the case of uniform, but arbitrarily strong, electric field, the approach based on Airy-functions has proven to be quite succesful (Bertoncini *et al.*, 1989, 1990; Bertoncini and Jauho, 1991, 1992). The research area has not yet reached a consensus: it is not clear whether the *intra-collisional field effect* (Barker, 1973; Barker and Ferry, 1979), around which much of the above quoted papers have focused, plays a significant role in real semiconductor microstructures (see, for example, Abdolsalami and Khan, 1990; Lipavsky *et al.*, 1991).

Nonequlibirum Green's Functions and Tunneling

In the context of tunneling problems the time-independent nonequilibrium formalism works as follows. In the remote past the contacts (i.e. the left and right lead) and the central region are decoupled, and each region is in thermal equilibrium. The equilibrium distribution functions for the three regions are characterized by their respective chemical potentials; these do not have to coincide nor are the differences between the chemical potentials necessarily small. The couplings between the different regions are then established and treated as perturbations via the standard techniques of perturbation theory, albeit on the two-branch time contour. It is important to notice that the couplings do not have to be small, e.g. with respect to level spacings or k_BT, and typically must be treated to all orders.

The time-dependent case can be treated similarly. Before the couplings between the various regions are turned on, the single-particle energies acquire rigid time-dependent shifts, which, in the case of the non-interacting contacts, translate into extra phase factors for the propagators (but not in changes in occupations). The perturbation theory with respect to the couplings has the same diagrammatic structure as in the stationary case. The calculations, of course, become more complicated because of the broken time translational invariance.

Model Hamiltonian

We split the total Hamiltonian in three pieces: $H=H_c+H_T+H_{cen}$, where H_c describes the contacts, H_T is the tunneling coupling, and H_{cen} models the interacting central region, respectively. Below we discuss each of these terms.

Contacts. Guided by the typical experimental geometry in which the leads rapidly broaden into metallic contacts, we view electrons in the leads as non-interacting except for an overall self-consistent potential. Physically, applying a time-dependent bias (electrostatic-potential difference) between the source and drain contacts means that the single-particle energies become time dependent: $\varepsilon_{k\alpha}^0 \to \varepsilon_{k\alpha}(t)=\varepsilon_{k\alpha}^0+\Delta_\alpha(t)$ (here α lables the channel in the left (L) or right (R) lead). The occupation of each state remains unchanged, and is determined by an equilibrium distribution established in the distant past. Thus the contact Hamiltonian is

$$H_c=\sum_{k\alpha\in L,R}\varepsilon_{k\alpha}(t)c_{k\alpha}^+c_{k\alpha}\ , \tag{11}$$

and the exact time-dependent Green's functions in the leads for the uncoupled system are

$$g_{k\alpha}^<(t,t')=if(\varepsilon_{k\alpha}^0)\exp\left[-i\int_{t'}^{t}dt_1\varepsilon_{k\alpha}(t_1)\right], \tag{12}$$

$$g_{k\alpha}^{r,a}(t,t')=\mu i\theta(\pm t\,\mu\,t')\exp\left[-i\int_{t'}^{t}dt_1\varepsilon_{k\alpha}(t_1)\right]. \tag{13}$$

Our model differs from the choice made in a recent study of Chen and Ting (1991): these authors allow the electrochemical potentials in the distribution functions to vary with time. In our view, this may result in an unphysical pile-up (or depletion) of charges: in real devices it is only the relatively small number of electrons in the accumulation/depeletion layer that is time-dependent. Further, the model studied by Chen and Ting does not display any of the interference phenomena predicted by our phase-conserving model.

Coupling Between Leads and Central Region. These couplings can be modified with time-dependent gate voltages, as is the case in single-electron pumps. The precise functional form of the time-dependence is determined by the detailed geometry and by the self-consistent response of charge in the contacts to external driving. We assume that these parameters are known (see also discussion below), and simply write

$$H_T = \sum_{k\alpha n} [\, V_{k\alpha,n}(t) c^+_{k\alpha} d_n + h.c. \,] \,. \tag{14}$$

Here the d's form a complete orthonormal set of single-electron creation and annihilation operators in the interacting region.

The Central Region. The form chosen for the central region Hamiltonian depends on geometry and on the physical behavior being investigated. Our results relating the current to local properties, such as densities of states and Green's functions, are valid generally. To make the results more concrete, we will discuss two particular examples in detail. In the first, the central region is taken to consist of noninteracting, but time-dependent levels,

$$H_{cen} = \sum_n \varepsilon_n(t) d^+_n d_n \,. \tag{15}$$

This choice represents a simple model for time-dependent resonant tunneling. We have derived general formulae valid for an arbitrary number of levels, but below we analyze, for simplicity, only the case for a single level in detail. The second example we will discuss is resonant tunneling with electron-phonon interaction,

$$H^{el-ph}_{cen} = \varepsilon_0 d^+ d + d^+ d \sum_q M_q \left[a^+_q + a_{-q} \right] . \tag{16}$$

In the above, the first term represents a single site, while the second term represents interaction of an electron on the site with phonons. The full Hamiltonian of the system must also include the free-phonon contribution. This example, while not exactly solvable, is helpful to show how interactions influence the current. Furthermore, we can compare to previous time-independent results (Wingreen *et al.*, 1989; Glazman and Shekter, 1987) using (16) to demonstrate the power of the present formalism. Another model, which is very important in connection with interacting mesoscopic transport, is the Anderson-impurity model,

$$H_A = \sum_\sigma \varepsilon_\sigma d^+_\sigma d_\sigma + U n_\uparrow n_\downarrow \,. \tag{17}$$

The second term accounts for the Coulomb repulsion of two electrons with opposite spins residing in the same site. This model can also be analyzed with the methods described below, but the scope of this review does not allow an in-depth discussion of the rich physics contained in (17), and we refer the reader to the rapidly growing literature on the topic, which includes both experimental and theoretical work (see, for example, Hershfield *et al.*, 1991; Meir and Wingreen, 1992; Ng, 1993; Yeyati *et al.*, 1993; Ralph and Buhrman, 1994).

Validity of the Model. The time-dependent problem has to be formulated carefully, particularly with respect to the leads. It is essential to a Landauer type of approach, that the electrons in the leads be non-interacting. In practice, however, the electrons in the leads near the mesoscopic region contribute to the self-consistent potential. As the model presented above suggests, we approach this problem by dividing the transport physics in two steps: (i) the self-consistent determination of charge pile-up and depletion in the contacts, the resulting barrier heights, and single-particle energies in the interacting region, and (ii) transport in a system defined by these self-consistent parameters. A similar approach has recently been developed by Büttiker and co-workers (1993a,b,c). Step (i) requires a capacitance calculation for each specific geometry, and we do not address it in this paper. Instead, as our model indicates, we assume the results of (i) as time-dependent input parameters and give a full treatment of the transport through the mesoscopic region.

When relating our results to actual experiments some care must be exercised. Specifically, we calculate only the current flowing into the mesoscopic region, while the total time-dependent current measured in the contacts includes contributions from charge flowing in and out of accumulation and depletion regions in the leads. In the *time-averaged* (dc) current, however, these capacitive contributions vanish, and the corresponding time-averaged formulae, to be given below, are directly relevant to experiment. It should be noted, though, that these capacitive currents may influence the effective time-dependent parameters in step (i) above.

Let us next estimate the frequency limits that restrict the validity of our approach. Two criteria must be satisfied. First, the driving frequency must be sufficiently slow that the applied bias is dropped entirely across the tunneling structure. When a bias is applied to a sample, the electric field in the leads can only be screened if the driving frequency is smaller than the plasma frequency, which is tens of THz in typical doped semiconductor samples. For signals slower than this, the bias is established entirely across the tunneling structure by accumulation and depletion of charge near the barriers. The unscreened Coulomb interaction between net excess charge is quite strong, and hence the bias across a tunneling structure is caused by a relatively small excess of charge in accumulation and depletion layers.

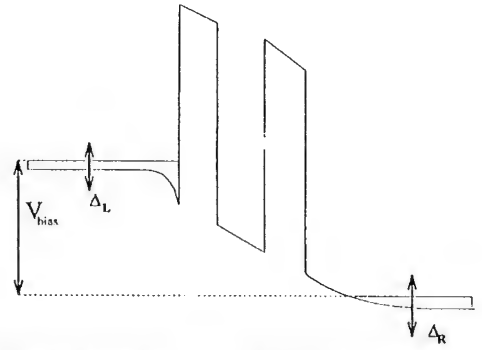


Figure 2. Sketch of charge distribution in a three dimensional resonant-tunneling device under dc-bias $V_{bias} = \mu_L - \mu_R$ with a time-modulation of amplitude $\Delta_{L/R}$ superposed on the leads. As argued in the text, only a tiny fraction of charge carriers participates in setting up the voltage drop across the structure.

The second frequency limit on our apporach is that the build-up of electrons required for the formation of the accumulation and depletion layers must not significantly disrupt the coherent transport of electrons incident from the leads. One way to quantify this is to ask - what is the probability that an electron incident from the leads participates in the build-up of charge associated with a time-dependent bias? This probability will be the ratio of the net current density flowing into the accumulation region to the total incident flux of electrons. For a three-dimensional double-barrier resonant-tunneling structure (see Fig. 2) the ac-current charging the accumulation layer is $I^{rms}_{acc} = 2\pi\upsilon C V^{rms} / A$, where υ is the driving

frequency, C is the capacitance, V^{ms} is the applied bias, and A is the area. In comparison, the total incident flux is $I_{inc} = 3/8 e n v_F$. Using the parameters appropriate for a typical experiment (Brown et al., 1991), we find that up to 10 THz the probability of an electron participating in the charge build-up is only 1%. Summarizing, these estimates indicate that our approach should be accurate up to frequencies of tens of THz, which are large by present experimental standards, and consequently the analysis presented should be valid for most experimental situtations.

Time-dependent Current

The current from the left contact through the left barrier to the central region can be calculated from the time evolution of the occupation number operator of the left contact:

$$J_L(t) = -e\langle \dot{N}_L \rangle = -\frac{ie}{\eta}\langle [H, N_L] \rangle , \tag{18}$$

where N_L is the occupation number operator for the left contact. After lengthy calculations (Wingreen *et al.*, 1993; Jauho *et al.*, 1994) it is possible to transform (18) into an expression which formally involves only the central region Green's function

$$J_L(t) = -\frac{2e}{\eta}\int_{-\infty}^{t} dt_1 \int \frac{d\varepsilon}{2\pi} \mathrm{Im}\, Tr\left\{ \exp\left[-i\varepsilon(t_1 - t)\right] \Gamma^L(\varepsilon, t_1, t)\left[G^<(t, t_1) + f_L(\varepsilon) G^r(t, t_1)\right]\right\}. \tag{19}$$

Here the Green's functions are matrices in the central region quantum numbers m,n, and the generalized level-width function Γ is defined by

$$\left[\Gamma^L(\varepsilon, t_1, t)\right]_{mn} = 2\pi \sum_{\alpha \in L} \rho_\alpha(\varepsilon) V_{\alpha,n}(\varepsilon, t) V^*_{\alpha,m}(\varepsilon, t_1) \exp\left[i\int_{t_1}^{t} dt_2 \Delta_\alpha(\varepsilon, t_2)\right] , \tag{20}$$

where we introduced the densities of states in the leads: $\rho_\alpha(\varepsilon)$. An analogous formula applies for the current flowing into the central region through the right barrier. Equation (19) is a very general formula, and it is valid for a wide range of applications. It expresses the current in terms of *local* quantities: Green's functions of the central region. The first term in (19), which is proportional to $G^<$, suggests an interpretation as the out-tunneling rate (recalling $\mathrm{Im} G^<(t,t) = N(t)$). Likewise, the second term, which is proportional to the occupation in the leads and to the density of states in the central region, can be asociated to the in-tunneling rate. It is important to bear in mind that *all* Green's functions in (19) are to be calculated in the presence of tunneling. Thus, in general, $G^<$ will depend on the occupation in the leads. Furthermore, in the presence of interactions, the retarded Green's function will depend on the central region occupation. Consequently, the current can be a non-linear function of the occupation factors. This issue has recently been discussed by other authors (Lake et. al (1993)).

Time-independent case

In the time-independent case the level-width function simplifies, and (19) can be written in a compact form:

$$J = \frac{ie}{2\eta}\int \frac{d\varepsilon}{2\pi} Tr\left\{ \left[\Gamma^L(\varepsilon) - \Gamma^R(\varepsilon)\right] G^<(\varepsilon) \right.$$

$$\left. + \left[f_L(\varepsilon)\Gamma^L(\varepsilon) - f_R(\varepsilon)\Gamma^R(\varepsilon)\right]\left[G^r(\varepsilon) - G^a(\varepsilon)\right] \right\}. \tag{21}$$

In writing (21) we recalled that in steady state the current is uniform, and can thus be symmetrized with respect to left and right contributions. Even further simplifcation can be achieved if the level-width functions are proprotional to each other: $\Gamma^L(\varepsilon) = \lambda \Gamma^R(\varepsilon)$. In this case we find

$$J = \frac{ie}{\eta}\int \frac{d\varepsilon}{2\pi}\left[f_L(\varepsilon) - f_R(\varepsilon)\right] Tr\left\{ \frac{\Gamma^L(\varepsilon)\Gamma^R(\varepsilon)}{\Gamma^L(\varepsilon) + \Gamma^R(\varepsilon)}\left(G^r(\varepsilon) - G^a(\varepsilon)\right)\right\}. \tag{22}$$

The difference between the retarded and advanced Green's functions is essentially the density of states. Despite of the apparent similarity of (22) to the Landauer formula, it is important to bear in mind that in general there is no immediate connection betwen the quantity in curly brackets, and the transmission coefficient. In particular, when inelastic scattering is present, we do not believe that such a connection exists. For the case of a noninteracing central region a connection with the transmission coefficient *can* be established. Further, it can be shown that the *time-averaged* current such a connection exists. Starting from (19), one can show (Jauho et al., 1994) that

$$\langle J(t)\rangle = -\frac{2e}{\eta}\int \frac{d\varepsilon}{2\pi}\left[f_L(\varepsilon) - f_R(\varepsilon)\right] \mathrm{Im}\, Tr\left\{ \frac{\Gamma^L(\varepsilon)\Gamma^R(\varepsilon)}{\Gamma^L(\varepsilon) + \Gamma^R(\varepsilon)}\langle A(\varepsilon, t)\rangle\right\}. \tag{23}$$

Here the time-average of a time-dependent object $F(t)$ is defined by

$$\langle F(t)\rangle = \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2} dt F(t) , \tag{24}$$

and

$$A(\varepsilon, t) = \int dt_1 G^r(t, t_1) \exp\left[i\varepsilon(t - t_1) + i\int_{t_1}^{t} dt_2 \Delta(t_2)\right] . \tag{25}$$

To obtain (23) one must assume that the time-variations in the left and right contact are identical. One often encounters situations where the energy-dependence of the level width functions is not essential; then the time-dependences in the contacts do not have to be related. The object A is central to much of the analysis to follow, and below we shall see explicit examples of this function.

Non-interacting resonant level model; time-independent case

For this important, and often encountered case, a simple analytic formula for the current can be derived:

$$J = \frac{e}{\eta}\int \frac{d\varepsilon}{2\pi} \frac{\Gamma^L(\varepsilon)\Gamma^R(\varepsilon)}{\left[\varepsilon - \varepsilon_0 - \Lambda(\varepsilon)\right]^2 + \left[\Gamma(\varepsilon)/2\right]^2}\left[f_L(\varepsilon) - f_R(\varepsilon)\right] . \tag{26}$$

Here $\Gamma \equiv \Gamma^L + \Gamma^R$ is the imaginary part of the retarded self-energy, and Λ is the corresponding real part:

$$\Sigma^{r,a}(\varepsilon) = \sum_{k\alpha \in L,R} \frac{|V_{k\alpha}|^2}{\varepsilon - \varepsilon_{k\alpha} \pm i\eta} = \Lambda(\varepsilon) \mu \frac{i}{2}\Gamma(\varepsilon) . \tag{27}$$

The derivation leading to (26) has made no asumptions about the energy dependence of the coupling to the leads. The factor multiplying the difference of Fermi functions is the elastic transmisison coefficient. It is important to understand the difference between (26) and (22): (22) gives the current for a *fully interacting system* , and the evaluation of the retarded and advanced Green's functions requires a consideration of interactions (e.g., electron-electron, electron-phonon, and spin-flip) in addition to tunneling back and forth to the contacts. To further illustrate the differences between (22) and (26), suppose now that the Green's function for the interacting central region can be solved:

$$G^{r,a}(\varepsilon) = \frac{1}{\varepsilon - \varepsilon_0 - \lambda(\varepsilon) \pm \frac{i}{2}\gamma(\varepsilon)} , \tag{28}$$

where λ and $\gamma/2$ are the real and imaginary parts of the self-energy (including interactions *and* tunneling. Then the interacting results for proportionate coupling (22) becomes

$$J = \frac{e}{\eta}\int \frac{d\varepsilon}{2\pi}[f_L(\varepsilon) - f_R(\varepsilon)] \frac{\Gamma^L(\varepsilon)\Gamma^R(\varepsilon)}{\Gamma^L(\varepsilon) + \Gamma^R(\varepsilon)} \frac{\gamma(\varepsilon)}{[\varepsilon - \varepsilon_0 - \lambda(\varepsilon)]^2 + [\gamma(\varepsilon)/2]^2} . \tag{29}$$

This expression coincides with the non-interacting result when the interactions are turned off, i.e. $\lambda \to \Lambda$ and $\gamma \to \Gamma$. In a phenomenological model, where the total level width is expressed as a sum of elastic and inelastic widths, $\gamma \to \gamma_e + \gamma_i$, one recovers the results of Jonson and Grincwajg (1987), and Weil and Vinter (1987).

Non-interacting resonant level model; wide-band limit

The energy-dependence of the level-width function $\Gamma(\varepsilon)$ is often not essential (in particular, this is the case when transport is dominated by states close to the Fermi level), and one may ignore it. Technically, the *wide-band limit* consists of the following steps: i) the level shift $\Lambda(\varepsilon)$ is ignored, ii) the line-widths are assumed to be energy independent constants, $\sum_{\alpha \in L,R} \Gamma_\alpha = \Gamma^{L/R}$, and iii) a single time-dependence, $\Delta_{L/R}(t)$, is allowed for the energies in each lead. These assumptions allow one to derive the following expression for the time-dependent current flowing into the central region from the left(right) contact:

$$J_{L/R}(t) = J^{out}_{L/R}(t) + J^{in}_{L/R}(t) , \tag{30}$$

where

$$J^{out}_{L/R}(t) = -\frac{e}{\eta}\Gamma^{L/R} N(t) , \tag{31}$$

$$J^{in}_{L/R}(t) = -\frac{e}{\eta}\Gamma^{L/R}\int \frac{d\varepsilon}{\pi} f_{L/R}(\varepsilon)\,\mathrm{Im}\{A_{L/R}(\varepsilon,t)\} . \tag{32}$$

Here the occupation $N(t)$ is given by

$$N(t) = \sum_{L/R}\Gamma^{L/R}\int \frac{d\varepsilon}{2\pi} f_{L/R}(\varepsilon)\left|A_{L/R}(\varepsilon,t)\right|^2 , \tag{33}$$

and the function A was defined in (25). We shall next consider two special cases, which are relevant to experimental situations.

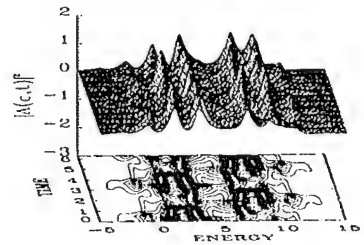


Figure 3. $|A(\varepsilon,t)|^2$ **as a function of time for harmonic modulation for a symmetric structure,** $\Gamma^L = \Gamma^R = \Gamma/2$. **The unit for the time-exis is** η/Γ **and all energies are measured in units of** Γ, **with the values** $\mu_L = 10, \mu_R = 0, \varepsilon_0 = 5, \Delta_0 = 5, \Delta_L = 10$ **and** $\Delta_R = 0$. **The modulation frequency is** $\omega = 2\Gamma/\eta$.

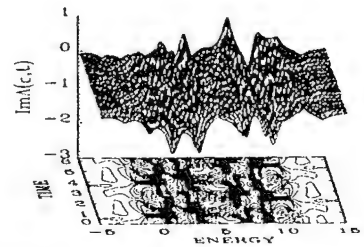


Figure 4. The time-dependence of $\mathrm{Im}\,A(\varepsilon,t)$ **for the case shown in Fig. 5.**

Response to harmonic modulation. Harmonic time modulation is probably the most commonly encountered example of time dependence. Here we give results for the case when the contact and site energy levels vary as

$$\Delta_{L/R,0}(t) = \Delta_{L/R,0}\cos(\omega t) . \tag{34}$$

The function $A(t)$ can be evaluated analytically, the resulting cumbersome expression can be found in Jauho et. al. (1994). In Figures 3 and 4 we show square modulus and imaginary part of A, respectively, as a function of time. An examination of these figures helps to understand the complicated time dependence of the current, to be discussed below. In Figures 3 and 4 the three-dimensional plot (top part of figure) is projected down on a plane

to yield a contour plot in order to help to visualize the time dependence. As expected, the time variation is periodic with period $T = 2\pi/\omega$. The time dependence is strikingly complex. The most easily recognized features are the maxima in the plot for $|A(\varepsilon, t)|$; these are related to photon side-bands occuring at $\varepsilon = \varepsilon_0 \pm k\omega$.

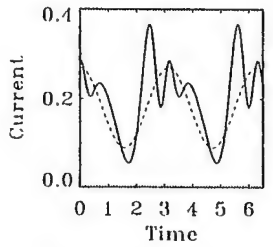


Figure 5. The time-dependent current $J(t)$ for harmonic modulation corresponding to the parameters of Figures 3 and 4. The dc bias is defined via $\mu_L = 10$ and $\mu_R = 0$, respectively. The dotted line shows (not drawn to scale) the time dependence of the drive signal. The temperature is $k_B T = 0.1\Gamma$.

Evaluation of the current requires a numerical integration of (31)-(33), and the results are shown in Fig. 5. We also display the drive voltage as a broken line. Bearing in mind the complex time dependence of $|A|^2$ and ImA, which determine the out- and in-currents, respectively, it is not surprising that the current displays a non-adiabatic time dependence. The basic physical mechanism underlying the secondary maxima and minima in the current is the line-up of a photon-assisted resonant-tunneling peak with the contact chemical postentials. The rapid time variations are due to J^{in} (or equivalently, due to ImA): the out-current J^{out} is determined by the occupation $N(t)$, and hence varies only on a time-scale Γ/η, which is the time scale for charge density changes.

We next consider the time-average of the current. For the case of a harmonic time dependence, we find

$$\langle \mathrm{Im}\, A_{L/R}(\varepsilon, t)\rangle = -\frac{\Gamma}{2}\sum_{k=-\infty}^{\infty} J_k^2\left(\frac{\Delta_0 - \Delta_{L/R}}{\omega}\right)\frac{1}{(\varepsilon - \varepsilon_0 - k\omega)^2 + (\Gamma/2)^2}\,. \qquad (35)$$

Figure 6 shows the resulting time-averaged current J_{dc}. A consequence of the complex harmonic structure of the time-dependent current is that for temperatures $k_B T < \eta\omega$ the average current oscillates as a function of period $2\pi/\omega$. The oscillation can be understood by examining the general expression for average current (23) together with (35): whenever a photon-asisted peak in the effective density of states, occuring at $\varepsilon = \varepsilon \pm k\omega$ in the time-averaged density of states $\langle \mathrm{Im}\, A_{L/R}\rangle$, moves in or out of the allowed energy range, determined by the difference of the contact occupation factors, a maximum (or minimum) in the average current results.

Response to steplike modulation. We give results for the case when the central level changes abruptly at $t = t_0$: $\varepsilon_0 \to \varepsilon_0 + \Delta$. One finds for $t > t_0$ from (25)

$$A(\varepsilon, t) = \frac{1}{\varepsilon - \varepsilon_0 + i\Gamma/2}\left[1 + \Delta\,\frac{1 - \exp\left(i(\varepsilon - \varepsilon_0 - \Delta + i\Gamma/2)(t - t_0)\right)}{\varepsilon - \varepsilon_0 - \Delta + i\Gamma/2}\right]. \qquad (36)$$

This result is easily genergalized to a pulse of duration s, and numerical results are discussed below. Just like in the case of harmonic modulation, it is instructive to study the time dependence of $|A^2|$ and ImA: these are shown in Figures 7 and 8. The observed time dependence is less complex than in the harmonic case. Nevertheless, the resulting current, which we have computed for a pulse of duration s, and display in Figure 9, shows an interesting ringing behavior. The ringing is again due to the movement of the sidebands of ImA through the contact Fermi energies.

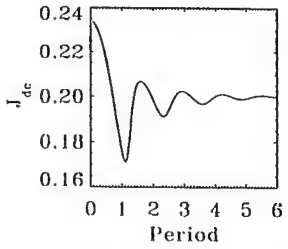


Figure 6. Time averaged current J_{dc} as function of the ac oscillation period $2\pi/\omega$. The dc amplitudes are the same as in those in Fig. 5.

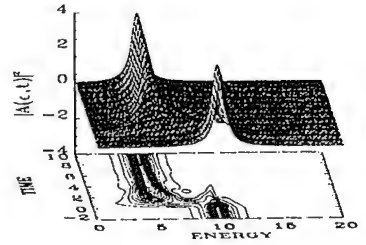


Figure 7. $|A(\varepsilon, t)|^2$ as a function of time for step-like modulation. At t=0 the resonant-level energy ε_0 suddenly decreases by 5Γ.

Due to the experimental caveats discussed above, the ringing shown in Fig. 9 may be masked by capacitive effects not included in the present work. However, the ringing should be observable in the time-averaged current by applying a series of pulses such as that of Figure 8, and then varying the pulse duration. In Fig. 10, the derivative of the dc current with respect to pulse length is plotted, normalized by the repeat time τ between the pulses. Fore pulse lengths s or the order of the resonance lifetime η/Γ, the derivative of the dc current mimics closely the time-dependent current following the pulse, and, likewise, asymptotes to the steady-state current at the new voltage.

Linear response. For circuit modeling purposes it would often be desirable to replace the mesoscopic device with a conventional circuit element, with an associated complex impedance $Z(\omega)$, or admittance $Y(\omega)$. Our results for the nonlinear time-dependent current form a very parctical starting point for such a calculation. For the non-

interacting case, the current is determined by $A(\varepsilon,t)$, and all one has to do is to linearize A with respect to the drive signal, i.e $\Delta - \Delta_{L/R}$. It is important to notice that one does not have to linearize with respect to the chemical potential difference, and thus our results are not limited to the equilibrium case (no bias voltage) studied recently by Fu and Dudley (1993).

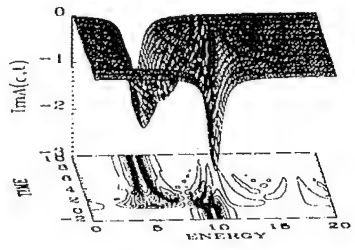


Figure 8. The time-dependence of ImA for the case in Fig. 7.

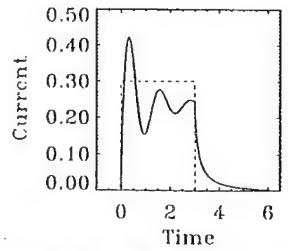


Figure 9. Time-dependent current $J(t)$ through a symmetric double-barrier tunneling structure in response to a rectangular bias pulse. Initially, the chemical potentials in left and right contacts and the resonant-level energy are all zero. At t=0, a bias pulse (dashed curve) suddenly increases energies in the left lead by $\Delta_L = 0$, and increases the resonant-level energy by $\Delta = 5$. At t=3, before the current has settled to a new steady value, the pulse ends and the current decays back to zero. The temperature is $k_B T = 0.1\Gamma$.

The linearization is straightforward though tedious procedure, and the explicit formulae are given in Jauho *et al.*, (1994), both for zero and finite temperatures. Here we show the numerical results for a model resonant tunneling diode. This model is a generalization of the wide-band limit studied above: in addition to the high-energy cut-off, provided by the chemical potential, we have introduced a low energy cut-off $D_{L/R}$ for the left and right contacts. The resulting current-voltage characteristic is shown in Fig. 11. We note that the strong increase in current, which is observed in experimental systems at very high voltages, is not present in our model: this is because we have ignored the bias-dependence of the barrier heights as well as any higher lying resonances. We show in Fig. 12 the resulting linear-response admittance for a symmetric structure. Several points are worth noticing. For dc bias $eV = 5$ (energies are given in units of Γ) the calculated admittance resembles qualitatively the results reported by Fu and Dudley for zero external bias, except that the change in sign for the imaginary part of Y is not seen. For zero external bias (not shown in the figure) our finite band-width model leads to an admittance, whose imaginary part changes sign, and thus the behavior found by Fu and Dudley cannot be ascribed to an artefact of their infinite band-width model. More interestingly, for dc bias in the NDR regime, the real part is negative for small frequencies. This simply reflects the fact that the device is operating under NDR bias conditions. At higher frequencies the real part becomes positive, thus indicating that further modeling along the lines sketched here may lead to important implications on the high-frequency response of resonant-tunneling structures.

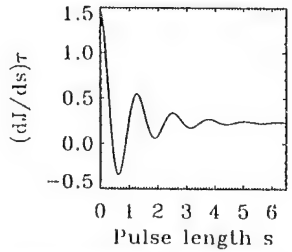


Figure 10. Derivative of the integrated dc current J_{dc} with respect to pulse duration s, normalized by the interval between pulses τ. For pulse durations much longer than the resonance lifetime τ/Γ, the derivative is just the steady-state current at the bias voltage, but for shorter pulses the ringing response of the current is evident.

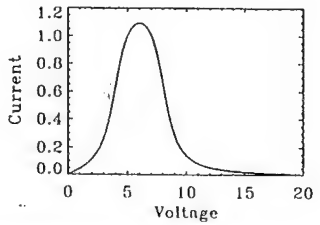


Figure 11. IV-characteristic for a model resonant-tunneling device (quantum dot).

Resonant tunneling with electron-phonon interactions

As the next application, we establish a connection to previous calculations on the effect of phonons on resonant tunneling (Wingreen *et al.*, 1989; Glazman and Shekter, 1987). For simplicity, we consider a single resonant level with energy-independent level widths. The expression for the current is now

$$J = \frac{e}{\eta}\frac{\Gamma^L\Gamma^R}{\Gamma^R+\Gamma^L}\int\frac{d\varepsilon}{2\pi}\left[f_L(\varepsilon)-f_R(\varepsilon)\right]\int_{-\infty}^{\infty}dt e^{i\omega}a(t), \tag{37}$$

where $a(t)$ is the interacting spectral density. In general, and exact evaluation of $a(t)$ is not possible: the hopping to and from the contacts results in an effective interaction between the electrons on the central site, and one would have to solve a true many-body problem.

However, if one ignores the Fermi sea, the retarded Green's function (and hence $a(t)$) can be calculated exactly (Mahan, 1990):

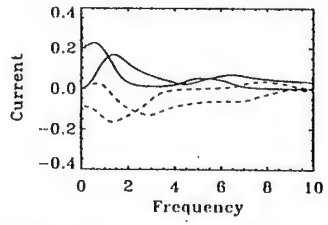


Figure 12 In-phase and out-of-phase components of the linear response current for two bias points, eV=5 (continuous line) and eV=10 (dashed line). The out-of-phase components (or, equivalently, imaginary parts) always tend to zero for vanishing frequency, while the in-phase component can have either a positive or negative zero-frequency limit depending on the dc bias.

$$G^r(t) = -i\theta(t)\exp\left[-it\left(\varepsilon_0 - \Delta\right) - \Phi(t) - \Gamma t/2\right], \tag{38}$$

where

$$\Delta = \sum_q \frac{M_q^2}{\omega_q}, \tag{39}$$

and

$$\Phi(t) = \sum_q \frac{M_q^2}{\omega_q^2}\left[N_q\left(1 - e^{i\omega_q t}\right) + \left(N_q + 1\right)\left(1 - e^{-i\omega_q t}\right)\right], \tag{40}$$

and the electron-phonon interaction is given by (16). When subsituted in the expression for the current, one recovers the results of Wingreen *et al.*, (1989), which were originally derived by analyzing the much more complex two-particle Green's function $G(\tau, s, t) = \theta(t)\theta(t)\langle d(\tau - s)d^+(\tau)d(t)d^+(0)\rangle$. The advantage of the method presented here is that one only needs the single-particle Green's function to use the interacting current formula. Other systematic approaches to the single-particle Green's function can therefore be directly applied to the current (e.g. pertubation theory in the tunneling Hamiltonian). We believe that there is yet much interesting physics to be discovered in interacting resonant tunneling systems, and as examples we mention two recent studies: Zhang *et al.*, (1994) have studied plasmon assisted resonant tunneling in a double barrier system, and the theoretical model used to analyze the experiments is essentially identical to the electron-phonon model discussed in this section, and Imam *et al.*, (1994) have considered the influence of electromagnetic flucuations on electron tunneling via one non-degenerate resonant level (this problem is relevant for electron transport thourgh quantum dots in the Coulomb blockade regime). The effect of fluctuations is taken into account by adding field-induced phases to the tunneling amplitudes, $V_\alpha \to V_\alpha e^{i\varphi_\alpha(t)}$, with $\varphi_\alpha(t) = eV_\alpha(t)$, where $\tilde{V}$ is the fluctuating voltage. The current is calculated with the interacting current formula given above; one has to also average over the fluctuations, and this can be effected once the effective impedances of the structure are known. It would be beyond the scope of these lectures to give a full description of this work; rather we illustrate the kind of calculations one has to perform by discussing a simpler model, where the central region energy level $\varepsilon_0(t)$ is assumed to be a Gaussian random variable. An example of a physical system where this kind of time dependence is relevant is tunneling through paramagnetic barriers (Rubo, 1993). The time-average is defined by specifying the correlators:

$$\langle \varepsilon_0(t) \rangle = 0 , \tag{41}$$

$$\langle \varepsilon_0(t_1)\varepsilon_0(t_2) \rangle = f(t_1 - t_2) . \tag{42}$$

All higher odd correlators vanish, while all even correlators are given as a sum of products of all pairwise correlators. In the calculation of $\langle \mathrm{Im}\, A(\varepsilon, t) \rangle$ for the average current one needs $\left\langle \exp\left[-i\int_{t-\tau}^{t} dt_1 \varepsilon_0(t_1)\right]\right\rangle$; this is evaluated by expanding the exponential functions, averaging term-by-term using (41-42), and resumming, with the result

$$\langle \mathrm{Im}\, A(\varepsilon, t) \rangle = \frac{1}{2}\int_{-\infty}^{\infty} d\tau e^{i\varepsilon\tau - \Gamma|\tau|/2} \exp\left[-\frac{1}{2}\int \frac{d\omega}{2\pi}\frac{f(\omega)}{\omega^2} 4\sin^2\left(\frac{\omega\tau}{2}\right)\right], \tag{43}$$

where $f(\omega)$ is the Fourier transform of $f(t)$. A particularly simple result emerges if one assumes a white-noise spectrum: $f(\omega) = F$; then the only effect of the fluctuations is to enhance the line-width: $\Gamma/2 \to \Gamma/2 + F$ (Weil and Vinter, 1987; Jonson and Grincwajg, 1987). A more complicated spectrum would change the Lorentzian line-shape to some other functional form.

COHERENT TIME-DEPENDENT TRANSPORT IN SUPERLATTICES

Review of the semiclassical theory

The most straightforward approach to study transport in semiconductor superlattices is to use the Boltzmann transport equation. As mentioned in the introduction, one of our goals is to study systems with several occupied minibands; these systems are clearly beyond the range of validity of the semiclassical Boltzmann equation [we recall that one of the basic assumptions underlying the derivation of the Boltzmann equations is that the band index n be constant (Ashcroft and Mermin, 1976)]. Nevertheless, the Boltzmannian description of a one-band system serves as a useful reference point, since, as we shall see, the quantum mechanical density-matrix description leads to expressions which bear a close resemblance to the Boltzmann equation description, and we can consequently analyze in detail the quantum mechanical correction terms. The Boltzmann equation for a one-dimensional superlattice, subjected to a time-dependent electric field $F(t)$, reads

$$\frac{\partial f(p,t)}{\partial t} + F(t)\frac{\partial f(p,t)}{\partial p} = -\frac{1}{\tau_\varepsilon}\left(f(p,t) - f^0\left(\varepsilon_p\right)\right) - \frac{1}{2\tau_p}\left(f(p,t) - f(-p,t)\right). \tag{44}$$

Here we consider two relaxation mechanisms: energy relaxation, characterized by the relaxation time τ_ε, which relaxes the nonequilibrium distribution function f towards the equilbirum distribution function f^0, and momentum relaxation, characterized by τ_p, which

drives f towards its angular average. We assume that the energy dispersion is given by a tight-binding form,

$$\varepsilon_p = \Delta_0 + \frac{\Delta_1}{2}\cos\left(\frac{pd}{\eta}\right), \tag{45}$$

where the $\Delta_{0,1}$ specify the center and width of the miniband, respectively, and d is the superlattice periodicity. As usual, the band velocity is evaluated via $v(p) = \partial\varepsilon/\partial p$. One can analyze (44) by taking its moments (Ignatov *et al.*, 1993); only two moments are required, and one obtains the following set of coupled differential equations:

$$\begin{pmatrix}\dot{x}\\ \dot{y}\end{pmatrix} + \begin{pmatrix}1/\tau_v & \Omega\\ -\Omega & 1/\tau_\varepsilon\end{pmatrix}\begin{pmatrix}x\\ y\end{pmatrix} = \frac{d}{\hbar}\begin{pmatrix}\Delta_0\Omega\\ \varepsilon_T/\tau_\varepsilon\end{pmatrix}. \tag{46}$$

Here we have introduced new variables via $x(t)=v(t)$, $y(t) = d\varepsilon(t)/\eta$, and $\Omega(t) = deE(t)/\eta$. We have also defined $1/\tau_v \equiv 1/\tau_p + 1/\tau_\varepsilon$, and the thermal energy $\varepsilon_T \equiv \int \varepsilon_p f^0(\varepsilon_p) dp$, which can be expressed in terms of Bessel functions, if the equilibrium distribution function is a Maxwell-Boltzmann distribution.

A number of results can be deduced from (46). For example, the *steady-state velocity* for a time-independent driving field is (Esaki and Tsu, 1970)

$$v = \frac{d}{\eta}\sqrt{\frac{\tau_v}{\tau_\varepsilon}}(\Delta_0 - \varepsilon_T)\frac{E/E_c}{1+(E/E_c)^2}, \tag{47}$$

where $E_c = \eta/(ed\sqrt{\tau_v\tau_\varepsilon})$. This famous relation, which leads to a negative differential resistance for $E > E_c$, is one of the reasons why superlattices have attracted significant attention as structures with potential device applications.

For a *collisionless* system (46) is readily solved with the result

$$v(t) = v_0 \sin\left[\int_0^t dt'\Omega(t') - \theta\right], \tag{48}$$

where v_0 and θ are constants depending on the initial conditions. In particular, for dc field one finds that the velocity oscillates with the period $\omega_B = edE/\eta$. These oscillations are the famous Bloch oscillations, whose experimental verification took more than sixty years after their prediction. If the driving field is a sum of a static and a harmonic piece, $E(t) = E_0 + E_1\cos\omega t$, one finds

$$v(t) = v_0\sum_n J_n\left(\frac{edE_1}{\eta\omega}\right)\sin\left[(\omega_B - n\omega)t - \theta\right]. \tag{49}$$

This result has a number of consequences: i) If the driving frequency is an integral fraction of the Bloch frequency, the system may support a dc current, and ii) In the absence of the dc field, the drift velocity vanishes, $\langle v(t)\rangle = 0$, if the expression $edE_1/\eta\omega$ is a root of J_0. This phenomenon is called 'dynamical localization', or band-collapse (Holthaus (1992)).

As a final example, we consider the case with no momentum relaxation ($1/\tau_p = 0$). Eliminating y from (46) results in the following differential equation:

$$\Omega\ddot{x} + (2\Omega a - \dot{\Omega})\dot{x} + (\Omega^3 + \Omega a^2 - \dot{\Omega})x = \frac{d}{dt}(\Omega p) + \Omega(ap - \Omega q). \tag{50}$$

Here we used the short-hand notation $a = 1/\tau_\varepsilon$, and p and q denote the inhomogenous terms in (46). This equation can be solved for a number of cases of interest, but it is presented here because later we shall make a connection to the quantum mechanical two-band case.

Tight-binding model for a two-band system

In order to study the effect of band-to-band tunneling on Bloch oscillations we need to find a simple, yet non-trivial model, which describes a two-band superlattice under the influence of an external electric field. We have chosen to use the model of Fukuyama *et al.* (1973):

$$H = \sum_n \left[(\Delta_{01} - nF_0d)a_n^+a_n + (\Delta_{02} - nF_0d)b_n^+b_n \right.$$
$$\left. -\frac{\Delta_1}{4}(a_{n+1}^+a_n + a_n^+a_{n+1}) + \frac{\Delta_2}{4}(b_{n+1}^+b_n + b_n^+b_{n+1}) - F_0R_{12}(a_n^+b_n + b_n^+a_n)\right]. \tag{51}$$

Here the first two terms give the field-dependent energies of the isolated levels '1' and '2', the next two terms allow hopping from site to site (and thus broaden the sharp isolated levels to bands '1' and '2'), and the last term describes the field-dependent coupling between the two bands. We recall the following properties of (51): i) $F_0 = 0$. The energy spectrum is given by two independent tight-binding bands: $\varepsilon_i(k) = \Delta_{0i} + \frac{\Delta_i}{2}(-1)^i\cos kd$, i=1,2. ii) Finite F_0. The energy spectrum consists of two interpenetrating Wannier-Stark-ladders: $\varepsilon_i = \Delta_{0i} - F_0r_i + nF_0d$, i=1,2. This result is the proof of existence of Wannier-Stark-ladders in two-band systmes (Fukuyama et al. (1973)). These authors did not calculate how the parameters r_i depend on the system parameters; we have performed such numerical calculations and an example of is shown in Fig. 13.

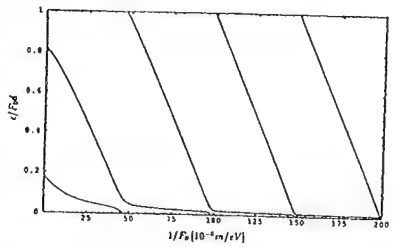


Figure 13. Energies of a tight-binding lattice under the influence of static electric field. The following parameters were used: $d = 10^{-8}m$, $\Delta_1 = 0.80\cdot10^{-2}$eV, $\Delta_2 = 0.92\cdot10^{-2}$, $\Delta_{02} - \Delta_{01} = 2.0\cdot10^{-2}$ eV, and $R_{12} = -16d/9\pi^2$.

Density-matrix equations for two bands

In what follows we assume that the applied electric field is uniform, but it may have arbitrary time-dependence. We use accelerated Bloch states as the basis set (Krieger and Iafrate, 1986,1987,1988), and since our aim is to study coherent transport we ignore scattering. With these assumptions, one can show that the density matrix $\rho_{1k}(t)$ obeys the following equation of motion:

$$\frac{\partial \rho_{1k}}{\partial t} = 2\,\mathrm{Re}\left\{ H_{12,k}(t)\int_0^t dt'\, H_{21,k}(t') \exp\left[-i\int_{t'} dt''\varepsilon_{21}(t'')\right]\left[\rho_{2k}(t') - \rho_{1k}(t')\right]\right\}. \tag{52}$$

$\rho_{2k}(t)$ satisfies a similar equation with the replacements $1 \leftrightarrow 2$. Here we have defined

$$\begin{aligned} H_{12,k}(t) &= -F(t)R_{12}(k(t)) \\ \varepsilon_{nn'}(t) &= \varepsilon_n(t) - \varepsilon_{n'}(t) \\ \varepsilon_n(t) &= \varepsilon_n^0(k(t)) - F(t)R_{12}(k(t)) \\ \hbar\dot{k} &= F(t)\,. \end{aligned} \tag{53}$$

It is important to note that the different k-components do not mix in the absence of scattering. In practical calculations this implies a vast simplification. However, one can include scattering within a relaxation-time approximation relatively easily.

From (52) it follows that the system has a constant of motion: $\frac{\partial}{\partial t}\left[\rho_{1k}(t) + \rho_{2k}(t)\right] = 0$, and consequently it is convenient to introduce a new function: $x(t) = \rho_{1k}(t) - \rho_{2k}(t)$. Calculations show that $x(t)$ satisfies the following equation of motion:

$$f(t)\ddot{x}(t) - \dot{f}(t)\dot{x}(t) + f^3(t)x(t) + f^2(t)\dot{g}(t)\int_0^t dt' \frac{\dot{g}(t')\dot{x}(t')}{f(t')} = 0\,. \tag{54}$$

Here we have used

$$\begin{aligned} f(k,t) &= 2F(t)R_{12}(k(t)) \\ g(k,t) &= \delta(k,t) - \int_0^t dt'\varepsilon_{21}(t') \\ \tan\delta(k,t) &= \frac{\mathrm{Im}\,R_{12}(k(t))}{\mathrm{Re}\,R_{12}(k(t))}\,. \end{aligned} \tag{55}$$

It is very interesting to note that the three first terms are *identical* to the first three terms in the semiclassical case (50), if one considers the collisionless case ($a=p=q=0$). Thus the quantum mechanical corrections are contained in the fourth term in (54). One must keep this term, if one wishes to study Zener tunneling.

Let us next study a special case of (54). We use the Kronig-Penney model for the superlattice, in which case $R_{12} = R_{21} = const.$, and the parameter δ vanishes. Further, we prepare the system in an initial equilibrium state, and turn the electric field on at $t = t_0$: $E(t) = E_0\theta(t - t_0)$. Then (54) simplifies to

$$\ddot{x} + f^2 x + \dot{g}\int_0^t dt'\, \dot{g}\dot{x} = 0\,. \tag{56}$$

It is relatively straightforward to solve this equation numerically, starting from a given initial state. Before turning to our numerical results, let us try to gain some qualitative understanding of what to expect. The initial state of the system is described by some wave-function, say $|\Psi(0)\rangle$, which is not an eigenstate after the field has been turned on. However, it can be expanded in terms of the eigenstates: $|\Psi(0)\rangle = \sum_n a_n\psi_{1,n} + b_n\psi_{2,n}$. The time-dependence of the density matrix is then given by

$$\rho(t) = |\Psi(t)\rangle\langle\Psi(t)| = \sum_{nm}\left[\, a_n^* e^{ic_n^{(1)}t}\psi_{1,n}^* + b_n^* e^{ic_n^{(2)}t}\psi_{2,n}^*\right]\left[a_m e^{-ic_m^{(1)}t}\psi_{1,m} + b_m e^{-ic_m^{(2)}t}\psi_{2,m}\right]. \tag{57}$$

We recall that the field-dependent eigenenergies are given by $\varepsilon_i = \Delta_{0i} - E_0 r_i + nE_0 d$, and thus the Fourier-transformed density matrix $\rho(\omega)$ is expected to have strong Fourier-components at

$$\omega = \begin{cases} mF_0 d \equiv m\omega_B \\ F_0(r_1 - r_2) + m\omega_B \\ F_0(r_2 - r_1) + m\omega_B \end{cases}. \tag{58}$$

Thus, solving $x(t)$ numerically, and working out its Fourier-transform, allows one to make a direct connection with the energy spectrum calculation shown in Fig. 12. First, Fig. 14 shows the time-dependence of $x(t)$ for three different field values. Thus, for certain field values, the time-dependence shows a sharp resonance, which will henceforth be called Zener-resonance. Physically, if $x(t)$ stays for all times in the proximity of unity, which is its initial value at $t=0$, there is no significant transfer from one miniband to the other. Correspondingly, if $x(t)$ drops to -1, there is a strong communication between the two bands. The situation is depicted in Fig. 15, where we show the field-dependence of the Zener resonance. We can understand origin of the resonances by studying Fig. 13: for certain values of the applied field the eigenenergies in the two wells are very close to each others (for example, at $1/E_0 \approx 97\cdot 10^{-8}$ m/eV, which corresponds to the middle panel in Fig. 14), and hence effectively at resonance, which leads to a very effective transfer between the two bands. In a simpler model, where the field-dependence of the energy-levels are ignored, one would find a monotonic (exponential) behavior for the tunneling rate.

Next, let us consider the Fourier components of $x(t)$. These are shown as dots in Fig. 16. (The relative darkness of the dots indicate the relative strength of the Fourier component.) The continuous lines are computed directly with (58). As is seen, the agreement between the two totally independent calculations is excellent. Of course, the density-matrix calculation contains much more information then the energy level determination (the intensities!); this information is required in the calculation of the current, and other experimentally relevant quantities. It is important to note that the k-value enters the calculation only through the initial value of $x(t)$, which should simplify the actual numerical calculations, which we hope to report in the near future.

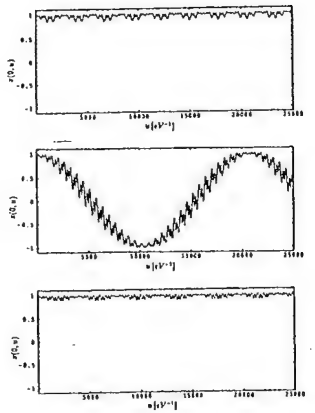


Figure 14. Time-dependence of the density-matrix for three different, but almost equal field values: $E_0 = 0.9 \cdot 10^6, 1.018 \cdot 10^6, 1.1 \cdot 10^6$ eV/m, respectively. The superlattice parameters are as in Fig. 13.

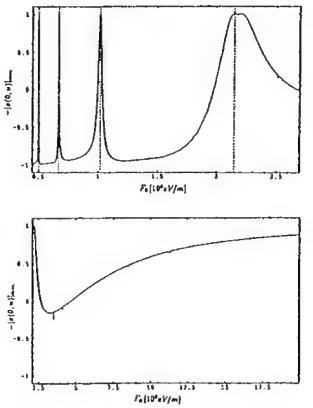


Figure 15. The negative of $x(t)$. The upper panel shows the low field regime, while the lower panel shows the high field regime. The dotted lines indicated the field strengths, where the Wannier-Stark levels are closest to each other.

Time-dependent external fields

As mentioned in the introduction, extremely strong time-dependent electric fields have become available with the development of free electron lasers. Thus the question arises: what is the effect of a strong time-dependent electric field on electric transport in superlattices? An interesting speculation was made by Holthaus (1992a, 1992b), who pointed out that at certain values of the external parameters a dynamical localization ('band-collapse') may occur. This result emerged already from the semiclassical analysis presented above, and here we address the same problem with the quantum mechanical density-matrix technique discussed in the previous section. In particular, we want to investigate, whether the band-collapse also persists in the two-band case.

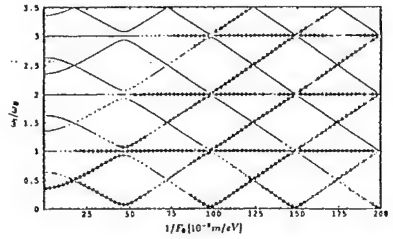


Figure 16. Fourier spectrum of $x(t)$. A dot corresponds to each significant peak in the Fourier spectrum; the darker the dot the more intense the spectral feature. The continuous lines are energy differences between the interpenetrating Stark Ladders.

One can make a strong parallel with independent electrons in a periodic potential on one hand, and independent electrons in a temporally periodic potential on the other hand. In the former case one can make use of Bloch's theorem, while in the time-dependent case one uses Flouquet's theorem to classify the allowed states and energies. Thus, given that the Hamiltonian is periodic in time, $H(t+T) = H(t)$, one can show that the wave-functions are of the form:

$$\psi_\varepsilon(t) = \exp(-i\varepsilon t) u_\varepsilon(t) \,, \tag{59}$$

where we introduced the *quasienergy* ε. The function u_ε has the same period as the Hamiltonian: $u_\varepsilon(t+T) = u_\varepsilon(t)$, and one can define a *quasi-Brillouin zone*, where the period $\Omega \equiv 2\pi/T$ plays the same role as the reciprocal lattice vector has for spatially periodic potentials. The quasienergy ε is defined only in modulo Ω.

We now want to repeat the analysis for the two-band tight-binding Hamiltonian (51) done in the previous sections, but now for a time-dependent electric field. As above, we divide the program in two steps: we first determine the quasienergy spectrum by a direct calculation, and next analyze the same problem with the density matrix formulation. We do not dwell on technical details, but present directly the obtained quasienergy spectrum in Fig. 17. We observe that the two-band model also displays a band-collapse, though not every time when non-interacting bands do. Thus, band-collapse seems to be a generic feature, and not an artefact of the previously studied one-band models.

The next step of the analysis is to solve the density-matrix equations with a time-dependent field. The situation is differs slightly from the static case: now the different k-values lead to solutions, which appear different, and to illustrate the point we display the results for two representative k-values in Fig. 18. As earlier, the dots indicate the strengths of the individual Fourier components. Just as in the static case, the two independent calculations are in excellent agreement.

SUMMARY

In the second part of these lectures we have discussed coherent transport properties of semiconductor superlattices, which are subjected to electric fields. We wish to emphasize the advantages of doing the analysis in time-domain: this may lead to significant new insights. As an example, we showed how the simple picture of Zener tunneling can acquire new aspects when viewed as a time-dependent phenomenon. The knowledge of the time-dependent density matrix gives new possibilities to analyze experiments probing Bloch oscillations, or Wannier-Stark ladders.

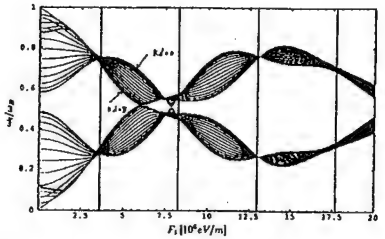


Figure 17. Quasienergy spectrum for the superlattice discussed in earlier examples. We display the field-dependence of 33 k-points, which are evenly distributed in the quasi-Brillouin zone. The vertical lines indicate the points where a band-collapse would occur for non-interacting minibands. $\hbar\omega_{ac}=1.5\cdot10^{-2}$ eV.

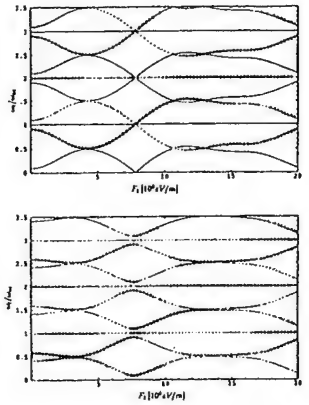


Figure 18. Fourier spectrum for a harmonically varying external field. Upper panel: kd=0; lower panel: kd=1.12. The continous lines are obtained from Figure 16.

It is clear that much remains to be done. We have not completed the current calculation, which requires the knowledge of all k-components of the density matrix. We have not included the effects of impurity scattering, even though we do not believe that this represents a major effort. If the charge carriers in the conduction band are created through optical excitation, one cannot ignore Coulombic effects between electrons and holes. All these open problems, and many others, invite future work in the area of time-dependent mesoscopic phenomena.

ACKNOWLEDGEMENTS

The work described in these notes has been performed in collaboration with N. S. Wingreen and Y. Meir (Green's function calculations), and with J. Rotvig and H. Smith (density-matrix calculations).

REFERENCES

Abdolsalami, F., and Khan, F., 1990, *Phys. Rev.* B41:3494.

Altshuler, B. L., Lee, P. A., and Webb, R. A, (eds.), 1991, "Mesoscopic Phenomena in Solids," Elsevier, Amsterdam.

Anda, E. V., and Flores, F., 1991, *J. Phys.: Condens. Matter* 3: 9087.

Ashcrof, N. W., and Mermin, N. D., 1976, "Solid State Physics," Holt, Rinehart and Winston, New York

Barker, J. R., 1973, *J. Phys. C* 6:2663.

Barker, J. R., and Ferry, D. K., 1979, *Phys. Rev. Lett.* 42:1779.

Bertoncini, R., Kriman, A. M., and Ferry, D. K., 1989, *Phys. Rev. B* 40:3371.

Bertoncini, R., Kriman, A. M., and Ferry, D. K., 1990, *Phys. Rev. B* 41:1390.

Bertoncini, R. and Jauho, A .P., 1991, *Phys. Rev. B* 44:3655.

Bertoncini, R. and Jauho, A. P., 1992, *Phys. Rev. Lett.* 68:2826.

Blandin, A., Nourtier, A., and Hone, D. W., 1976, *J. de Physique* 37:369.

Bloch, F., 1928, *Z. Physik* 52:555.

Brown, E. R. *et al.*, 1991, *Appl. Phys. Lett.* 58:2291.

Bruder, C., and Schoeller, H., 1994, *Phys. Rev. Lett.* 72:1076.

Büttiker, M., Prêtre, A., and Thomas, H., 1993a, *Phys. Rev. Lett.* 70:4114.

Büttiker, M., Thomas, H., and Prêtre, A., 1993b, *Physics Letters A* 180: 364

Büttiker, M., 1993c, *J. Phys. Condensed Matter* 5:9361.

Caroli, C., Combescot, R., Nozieres, P., and Saint-James, D., 1971a, *J. Phys. C.* 4:916.

Caroli, C., Combescot, R., Lederer, D., Nozieres, P., and Saint-James, D., 1971b, *J. Phys. C* 4:2589.

Caroli, C., Combescot, R., Nozieres, P., and Saint-James, D., 1972, *J. Phys. C* 5:21.

Chen, L. Y. and Su, Z. B., 1989, *Phys. Rev. B* 40:9309.

Chen, L. Y. and Ting, C. S., 1991, *Phys. Rev. B* 43:2097.

Davies, J. H., Hershfield, S., Hyldgaard, P., and Wilkins, J. W., 1993, *Phys. Rev. B* 47:4603.

Esaki, L. and Tsu, R., 1970, *IBM J. Res. Dev.* 14:61.

Fu, Y. and Dudley, S. C., 1993, *Phys. Rev. Lett.* 70:65.

Fukuyama, H., Bari, R.A., and Fogedby, H. C., 1973, *Phys. Rev. B* 8:5579.

Glazman, L. I. and Shekter, R. I., 1987, *Zh. Eksp. Teor. Fiz.* 94: 292 (1988, *Sov. Phys. JETP* 67:163).

Guimarães, P. S. S., Keay, B. J., Kaminski, J. P., Allen, S. J. Jr., Hopkins, P. F., Gossard, A. C., Florez, L. T., and Harbison, J. P., 1993, *Phys. Rev. Lett.* 70:3792.

Haug, H., and Jauho, A. P., 1994, "Quantum Kinetics in Transport and Optics of Semiconductors," to be published by Springer, Frankfurt.

Hershfield, S., and Ambegaokar, V., 1986, *Phys. Rev.* 34:2147.

Hershfield, S., Davies, J. H., and Wilkins, J. W., 1991, *Phys. Rev. Lett.* 67:3720.

Holthaus, M., 1992a, *Phys. Rev. Lett.* 69:351.

Holthaus, M., 1992b, *Z. Phys. B.- Condensed Matter* 89:251.

Holthaus, M., and Hone, D., 1993, *Phys. Rev. B* 47:6499.
Ignatov, A. A., Renk, K. F., and Dodin, E. P, 1993, *Phys. Rev. Lett.* 70:1996.
Imam, H. T., Ponomarenko, V. V., and Averin, D.V., 1994 (unpublished).
Jauho, A. P., and Wilkins, J. W., 1982, *Phys. Rev. Lett.* 49:762.
Jauho, A. P., and Wilkins, J. W., 1984, *Phys. Rev. B* 29:1919.
Jauho, A. P., 1989, *Solid State Electronics* 32:1265.
Jauho, A. P., 1991, *in* "Granular Nanoelectronics," Vol. 251 of Nato Advanced Study Institute, Series B: Physics, Eds. Ferry, D. K., Barker, J.R., and Jacoboni, C., Plenum Press, New York.
Jauho, A. P., Wingreen, N. S., and Meir, Y., 1994, Nordita Preprint 94/19 (to appear in *Phys. Rev. B)*
Jonson, M., and Grinczajg, A., 1987, *Appl. Phys. Lett.* 51:1729.
Kazarinov, R. F., and Suris, R. A., 1972, *Fiz. Tekh. Poluprov.* 6:148 (1972, *Sov. Phys. Semicond.* 6:120).
Khan, F.S., Davies, J.H., and Wilkins, J.W., 1987, *Phys. Rev. B* 36:2578.
Kirk, W. P., and Reed, M. A., (eds.), 1992, "Nanostructures and Mesoscopic Systems," Academic Press, San Diego.
Kadanoff, L. P., and Baym, G., 1962, "Quantum Statistical Mechanics," Benjamin, New York.
Keldysh, L.V., 1964, *Zh. Eksp. Teor. Fiz* 47:1515 (1965, *Sov. Phys. JETP* 20:1018).
Krieger, J. B., and Iafrate, G. J., 1986, *Phys. Rev. B* 33:5494.
Krieger, J. B., and Iafrate, G. J., 1987, *Phys. Rev. B* 35:9644.
Krieger, J. B., and Iafrate, G. J., 1988, *Phys. Rev. B* 38:6324.
Landauer, R., 1957, *IBM J. Res. Dev.* 1:233.
Landauer, R., 1970, *Philos. Mag.* 21:863.
Langreth, D. C., 1976, *in* "Linear and Nonlinear Electron Transport in Solids," Vol. 17 of Nato Advanced Study Institute, Series B: Physics, eds. Devreese, J.T. and Van Doren, V.E., Plenum, New York.
Langreth, D. C. and Nordlander, P., 1991, *Phys. Rev.* B 43:2541.
Lake, R., and Datta, S., 1992, *Phys. Rev. B* 45:6670.
Lake, R., Klimeck, G., Anantram, M., and Datta, S., 1993, *Phys. Rev. B* 48:15132.
Lee, P. A., and Ramakrishnan, T. V., 1985, *Rev. Mod. Phys.* 57:287.
Leo, K., Haring Bolivar, P., Brüggemann, F., Schwedler, R., and Köhler, K., 1992, *Solid State Commun.* 84:943.
Lipavsky, P., Spicka, V., and Velicky, B., 1986, *Phys. Rev. B* 34:6933.
Lipavsky, P., Khan, F. S., Abdolsalami, F., and Wilkins, J. S., 1991, *Phys. Rev. B* 43:4885.
Mahan, G. D., 1984, *Physics Reports* 110:231.
Mahan, G. D., 1987, *Physics Reports* 145:253.
Martin, P. C., and Schwinger, J., 1959, *J.Math. Phys.* 2:407.
Meir, Y. and Wingreen, N. S., 1992, *Phys. Rev. Lett.* 68:2512.
Meir, Y., Wingreen, N. S., and Lee, P. A., 1993, *Phys. Rev. Lett.* 70:2601.
Ng, T. K., 1993, *Phys. Rev. Lett.* 70:3635.
Pastawski, H. M., 1992, *Phys. Rev. B* 46:4053.
Ralph, D. C., and Buhrman, R. A., 1994, *Phys. Rev. Lett.* 21:3401.
Rammer, J., and Smith, H., 1986, *Rev. Mod. Phys.* 58:323.
Reggiani, L., Lugli, P., and Jauho, A. P., 1987, *Phys. Rev. B* 36:6602.
Rubo, Yu. G., 1993, *Zh. Eksp. Teor. Fiz.* 104:3536 (1993, *JETP* 77:685).
Runge, E., and Ehrenreich, H., 1992a, *Annals of Physics* 219:55.
Runge, E., and Ehrenreich, H., 1992b, *Phys. Rev.* B45:9145.
Schwinger, J., 1961, *Phys. Rev.* 115:1342.
Suhrke, M., and Wilke, S., 1992, *Phys. Rev. B* 46:2400.
Strinati, G., Castellani, C., and Di Castro, D., 1989, *Phys. Rev. B* 40:12237.
Vasko, F. T., 1983, *Fiz. Tverd. Tela* 25:20 (1983, *Sov. Phys. Solid. State* 25:10).
Waschke, Ch., Roskos, H. G., Schwedler, K., Leo, K., Kurz, H., and Köhler, K., 1993, *Phys. Rev. Lett.* 70:3319.
Weil, T., and Vinter, B., 1987, *Appl. Phys. Lett.* 50:1281.
Wingreen, N. S., Jacobsen, K. W., and Wilkins, J. W., 1989, *Phys. Rev.* B 40:11834.
Wingreen, N. S., Jauho, A. P., and Meir, Y., 1993, *Phys. Rev.* B 47:8487.
Yeyati, A. L., Martin-Rodero, A., and Flores, F., 1993, *Phys. Rev. Lett.* 71:2991.
Zener, C., 1934, *Proc. Roy. Soc. London Ser.* A145:523.
Zhang, C., Lerch, M. L. F., Martin, A. D., Simmonds, P. E., and Eaves, L., 1994, *Phys. Rev. Lett.* 72:3397.

DENSITY MATRIX SIMULATIONS OF SEMICONDUCTOR DEVICES

H. L. Grubin

Scientific Research Associates, Inc.
Glastonbury, Connecticut 06033 USA

INTRODUCTION

Since the pioneering work of Tsu and Esaki (1973), the experimental studies of Sollner *et al.* (1983) on double-barrier resonant tunneling devices, and the superlattice detector work of Levine *et al.* (1987), there has been growing interest in barrier/well devices and in the fundamental underpinnings of quantum device operation. Further, following the work Datta *et al.* (1986), there has also been rising interest in the basic physics accompanying the Aharonov-Bohm (1959) effect in heterostructures. Indeed, major advances in material technology has enabled device scientists to conjecture about new device structures that both test and illustrate basic fundamental quantum physics issues of few and many particle systems. For example the issue of *nonlocality* now finds its way into discussions of transport in quantum devices. Nonlocality in classical physics is illustrated by the coulomb interaction that decreases as the square of the distance between particles. In quantum mechanics there are additional interactions that do not necessarily drop off with distance and these are discussed below.

Another issue involves *dissipation*. Schrödinger's equation as traditionally used is dissipationless, and if all transport in subsystems were governed by Schrödinger's equation without interactions between the subsystems, all transport would be ballistic. Dissipation in quantum mechanics is treated by introducing additional systems, e.g., phonons, and allowing the additional system to cause a transition between states of the original system.

A third issue, specific to the treatment of electronic devices, is the *reservoir*. Traditionally, the examination of classical devices involves the specification of densities on the bounding surfaces, regarded as reservoirs. Such specification, which is assumed to remain valid under bias, often involves the concept of a quasi-Fermi level, in which the energy separation between the bottom of the conduction band and the Fermi level at the boundary remains unchanged.

Presently, our ability to incorporate these quantum mechanical issues to describe physical phenomena in ultra small devices and to propose quantum phase based devices has been evolutionary. Through a coupling of experiment, theory and numerical simulation we have been better able to understand how basic quantum mechanical processes affect device physics. But the 'goodness' of a description of quantum transport lies in the ability of the theory to explain the detailed experimental results obtained from such complex devices as, e.g., two terminal resonant tunneling diodes (RTD), quantum well superlattice detectors, and the more common heterostructure FETs. However, the complexity of the RTD and the puzzle associated with understanding its detailed operational principles has led Ferry (1989) to describe it as the *fruit fly* of quantum transport device theory. How good is the fruit fly analog.

Traditionally, transport in RTDs and other barrier structures has been analyzed through implementation of the formula (Tsu and Esaki, 1973):

$$J = [2e/(2\pi)^3]\int d\mathbf{k}v(\mathbf{k})[f_{FD}(E) - f_{FD}(E+e\phi)]|T(E,\phi)|^2 \ . \qquad (1)$$

It is the approximations associated with this formulae that provide the bounds of our understanding of transport in quantum structures. In (1), f_{FD} is the equilibrium Fermi-Dirac distribution function, $T(E,\phi)$ is the transmission coefficient obtained from solutions to the time independent Schrodinger equation, E is the energy of the tunneling particle and ϕ the applied potential. As discussed by Kluksdahl *et al.* (1989), a major criticism of this approach is that it requires knowledge of the distribution function at each side of the tunneling interface, rather than the bulk-like distribution far from the tunneling interface. Additionally the form of equation (1) also implies: (1) the use of equilibrium distribution functions to describe a biased state, when the biased resonant tunneling diode is in a non-equilibrium state; (2) the neglect of scattering, although scattering would be required to force a system to a state of equilibrium; and (3) the concept of a Fermi level, which clearly implies the presence of strong carrier-carrier interactions, particularly in the quantum well.

While the use of equation (1) has been successful in predicting negative conductance in RTD, its inadequacies in explaining experiment have been well documented. These include: *First*: the dc studies do not account for the peak-to-valley ratio of resonant tunneling devices. *Second*: the dc studies do not adequately treat dissipation. *Third*: the dc studies do not treat hysteresis in the current voltage characteristics, observed experimentally. *Fourth*: the dc treatment cannot predict how the devices will be used in applications. *Fifth*: the dc treatment cannot treat the time dependent nature of the boundary conditions that represent physical contacts.

The above studies suffer from lack of incorporating the feature basic to quantum mechanical phenomena: *all quantum mechanical devices are time-dependent*. Apart from dissipation, there are always reflections off boundaries, barriers, wells, imperfections and contacts. *What is needed are time dependent large signal numerical studies of quantum feature size devices*. This need has been discussed by Ravaioli *et al.* (1985) and Frensley (1990), and more recently Ferry and Grubin (1994). This approach emphasizes the details of transient behavior, the numbers of particles involved in device operation, the temporal duration under which the effective mass approximation is valid, the significance of the Fermi-golden rule, and other short time phenomena. As currently practiced, when scattering is present, or when time dependent fields are present and treated as perturbations, it is supposed that the perturbation does not modify the states of an unperturbed system; rather, the perturbed system, instead of remaining permanently in one of the unperturbed states is assumed to be continually changing from one to another, i.e., undergoing transitions from one state to another state. This approach is at the heart of those calculations employing the density matrix (See, e.g., Ehrenreich and Cohen, 1959; Goldstone and Gottfried, 1959;

Frensley, 1990; Krieger and Iafrate, 1987, 1989), those employing the Wigner distribution function (Wigner, 1932), and those employing Green's function techniques (Lake and Datta, 1992).

In addition to these fundamental approaches, there are also *derivative* procedures that enjoy wide spread use, both for the intuitive nature of the equations and because of the ease with which classical concepts emerge. These discussions include the quantum moment equations, see e.g., Iafrate *et al.* (1981), Stroscio (1986), and Grubin and Kreskovsky (1989).

In the discussion that follows the density matrix, Greens functions and quantum moment equations will be implemented to the study of quantum feature size devices. Further, we have found insight for multiparticle transport based upon concepts obtained through a recasting of the single particle Schrodinger equation. Adopting the approach of Bohm (Philippidis *et al.*, 1982; Bohm and Hiley, 1993) the single particle wave function is written in the form:

$$\psi = R\exp\left[\frac{iS}{\hbar}\right] , \tag{2}$$

subject to the condition that increasing the phase by 2π, does not change the wavefunction. This wave function, when inserted into Schrodinger's equation, results in two equations:

$$\frac{\partial S}{\partial t}+\frac{(\nabla S)^2}{2m}+V+Q=0 , \tag{3}$$

and

$$\frac{\partial R^2}{\partial t}+\nabla\cdot\left(R^2\frac{\nabla S}{m}\right)=0 , \tag{4}$$

where:

$$Q\equiv-\left(\frac{\hbar^2}{2m}\right)\frac{\nabla^2 R}{R} . \tag{5}$$

Equations (3)-(5) indicate that the Schrodinger wave represents a particle with a well defined position whose value is causally determined. The particle is never separate from the quantum force, $-\nabla Q$, that fundamentally affects it. The particle has an equation of motion:

$$m\frac{dv}{dt}=-\nabla(V+Q) , \tag{6}$$

which means that the forces acting on the particle consist of the classical force, $-\nabla V$, and the quantum force, $-\nabla Q$. It is important to note that the quantum potential is dependent on the shape of the real part of the wave function, rather than on its intensity; and does not necessarily fall off with distance. The quantum force is dependent on the momentum of the carrier through the continuity equation, but does not require a source term

The quantum potential is defined in terms of a single particle waverfunction. And if $S(\mathbf{r},t)=s(\mathbf{r},t)-Et$, where E, is a constant independent of position, then under zero current conditions, equation (3) is the real part of Schrodingers equation whose solutions, subject to a particular set of conditions, leads to a set of bound state eigenvalues. We will come back to this point over and over again, in the discussion that follows.

While the above discussion is for single particle wave functions we are interested in quantal and classical distribution functions, both representing an ensemble of particles. Our experience, developed from approximate representations of the Wigner distribution function (Wigner, 1932), indicates that the incorporation of the quantum potential for an ensemble of particles, where the amplitude R is replaced by the square root of the self-consistent density, $\rho(x)$, is a significant aid in interpreting much of the salient features of quantum transport in devices. The use of the quantum potential provides an alternative explanation for the peaking of the charge density at positions away from the interface of wide and narrowband gap structures; for real space transfer, for the potential distribution associated with a Schottky barrier, for density variations associated with variations in effective mass, and a host of additional features. To get to these points we must get through some mathematics, part of which is exact, and part approximate. We begin with the development of the single particle density matrix.

THE SINGLE PARTICLE DENSITY MATRIX

While the density matrix approach discussed below and the Wigner approach are mathematically equivalent, we have made the choice of the density matrix because the equation of motion readily submits to algorithms developed by T. R. Govindan; the use of which are extremely short computational times for steady state solutions. There are limitations to our treatment. The most important is that the equation of motion discussed below does not include anti-symmetric components and the density matrix has not been subject to anti-symmetrization (De Groot and Suttorp, 1972). We note that the application of the Wigner formulation to devices suffers from the same limitation. In some of the studies below, the inclusion of Fermi statistics is through the boundary conditions, as in the Wigner studies. It is not clear at the present time, whether this approach is sufficient for predicting device behavior.

The structures that we discuss fall under the category of *open structures* (Frensley, 1990), which can exchange particles with their surroundings, and which mathematically expresses this interaction in terms of boundary conditions. The phenomena we are interested in will be with systems that are far from equilibrium.

The density matrix is obtained from the density operator $\rho_{op}(t)$, which following Dirac notation (Dirac, 1958), is:

$$\rho_{op}(t)=\sum_i |i(t)>P(i)<i(t)| , \tag{7}$$

where $|i(t)>$ represents an eigenstate (Holland, 1993; a mixed state may be decomposed in an infinite number of ways, and so we cannot uniquely deduce from it the set of eigenstates and their respective weights). The time evolution of the density operator is obtained from the time evolution operator $U(t,t_0)$ which has the property $U(t,t_0)|i(t_0)>=|i(t)>$ (see Di-

race, 1958; section 27). The time evolution operator is unitary and the dependence of the density operator on previous times is given by:

$$\rho_{op}(t) = U(t,t_0)\,\rho_{op}(t_0)U(t,t_0)^{\dagger} \quad , \tag{8}$$

where the symbol '†' represents the adjoint. The time dependence of the density operator is governed by the time dependence of the time evolution operator, which is (Messiah, 1961):

$$i\hbar\frac{dU(t,t_0)}{dt} = H(t)U(t,t_0) \quad . \tag{9}$$

Assuming that the Hamiltonian *H(t)* is Hermetian:

$$i\hbar\frac{d\rho_{op}(t)}{dt} = \left[H(t),\,\rho_{op}(t)\right] \quad . \tag{10}$$

The density matrix in the coordinate representation is given by:

$$i\hbar\frac{\partial <x|\rho(t)|x'>}{\partial t} = \left\{-\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial x'^2}\right) + V(x) - V(x')\right\} <x|\rho(t)|x'> + i\hbar\left\{\frac{\partial <x|\rho|x'>}{\partial t}\right\}_{scattering} \quad . \tag{11}$$

Notice that we are ignoring any spatial variation in the effective mass, although we will deal with this later (Einevoll and Sham, 1994). The last term on the right hand side of equation (11) is a generic representation of scattering, which we treat below in a semiclassical manner. All of the quantum features associated with the devices below will arise from the streaming terms.

The density matrix is Hermetian, $\rho(x,x') \equiv <x|\rho_{op}|x'> = \rho(x',x)^{*}$. Additional quantities relevant for transport include the current density matrix:

$$j(x,x') = \frac{\hbar}{2mi}\left(\nabla_x - \nabla_{x'}\right)\rho(x,x') \quad , \tag{12}$$

and the energy density matrix:

$$E(x,x') = -\frac{\hbar^2}{8m}\left(\nabla_x - \nabla_{x'}\right)^2\rho(x,x') \quad . \tag{13}$$

The diagonal components of each represent the observables.

Equation (11) when coupled to Poisson's equation :

$$\nabla\cdot(\varepsilon\nabla)V = -e^2\left(\rho(x) - \rho_0(x)\right) \quad , \tag{14}$$

and *equations describing scattering* are the relevant equations for device transport. Note that while the above equations are for electrons, we will also discuss hole transport; the relevant modifications to the equation will be indicated.

The Liouville equation in the coordinate representation is a function of six variables plus time. The six variables represent a coordinate phase space whose relation to the standard phase space involving position and momenta may be assessed through application of the Weyl transformation (Weyl, 1927), which has been modified to include spin

$$\rho(x,x') = \left[\frac{2}{(2\pi)^3}\right]\int_{-\infty}^{+\infty} dk\, f_W\left[k, \frac{(x+x')}{2}\right]\exp\left[ik\cdot(x-x')\right] \quad . \tag{15}$$

To date the description of transport in devices via the density matrix has been confined to cases where the particles are free in two directions, which for specificity we take as the 'y' and 'z' directions. Further in the discussion below we will deal with diagonal components along the free directions, and treat the density matrix $\rho(x,x',y=y',z=z') \equiv \rho(x,x')$.

To determine the form of the density matrix, we can picture a situation in the absence of dissipation in which boundary conditions permit the separation of equations in two Schrodinger-type equations, with a solution that is the product of two wave functions. More generally we seek solutions of the type:

$$\rho(x,x',t) = \sum_i f_i\,\Psi_i^{*}(x',t)\Psi_i(x,t) \quad , \tag{16}$$

for which (15) is a special case. We now consider several examples.

EXAMPLES OF THE EQUILIBRIUM DENSITY MATRIX

For a Fermi-Dirac distribution function:

$$f_W(k,x) = \frac{1}{1+\exp\left[\left(E-E_F\right)/k_B T\right]} \quad , \tag{17}$$

and for parabolic bands the density matrix is:

$$\rho(x,x') = \left[\frac{N_C}{\Gamma(3/2)}\right]\left[\frac{\lambda}{(x-x')}\right]\int_0^{+\infty} d\mu\,\frac{\sin\left[\mu^{1/2}(x-x')/\lambda\right]}{1+\exp\left[\mu-\mu_n\right]} \quad . \tag{18}$$

Here,

$$Lim_{x\to x'}\,\rho(x,x') = N_c F_{1/2}(\mu_n),$$

$$F_{1/2}(\mu_n) = \left[\Gamma(3/2)\right]^{-1}\int_0^{+\infty}\frac{\mu^{1/2}d\mu}{1+\exp\left[\mu-\mu_n\right]},$$

$$\mu = (E-E_C)/k_B T, \quad \mu_n = (E_F - E_C)/k_B T,$$

$N_C = \Gamma(3/2)/(2\pi^{2}\lambda^{3})$ is the density of states and $\lambda^2 = \hbar^2/(2mk_B T)$ is *the square of the thermal deBroglie wavelength.*

There are two limiting cases that submit to analytical expression. In the high temperature limit, where Boltzmann statistics apply (the Boltzmann distribution arises when $\mu_n < -4$)

$$\rho(x,x') = N_c \exp\left[\mu_n - (x-x')^2 / 4\lambda^2\right] . \tag{19}$$

This distribution is Gaussian. For a material such as gallium arsenide, the thermal deBroglie wavelength at room temperature is 4.7 nm and $N_c = 4.4\times10^{23}/m^3$. For a nominal density of $10^{23}/m^3$, $\mu_n = -1.48$. In the low temperature limit, e.g., $T = 0$ K (March, 1987):

$$\rho(x,x') = \left[\frac{k_F^{\,3}}{\pi^2}\right]\frac{j_1\left[k_F(x-x')\right]}{k_F(x-x')} , \tag{20}$$

where $j_1(z)$ is a spherical Bessel function, $E_F = \hbar^2 k_F^2 / 2m$, and $k_F = [3\pi^2 N]^{1/3}$. In the limit as $z \Rightarrow 0$, $j_1(z) \Rightarrow z/3$. One of the earliest applications involving (20) was in a discussion by Bardeen (1936), where it was demonstrated that the electron density profile a distance 'z' from an infinite barrier was:

$$\rho(z) = N\begin{bmatrix} 1 - \dfrac{3}{2}\dfrac{j_1[2k_F z]}{k_F z} & for\ z>0 \\ 0 & otherwise \end{bmatrix} . \tag{21}$$

Figure 1 displays the density matrix corresponding to (20) for a density of $10^{23}/m^3$.

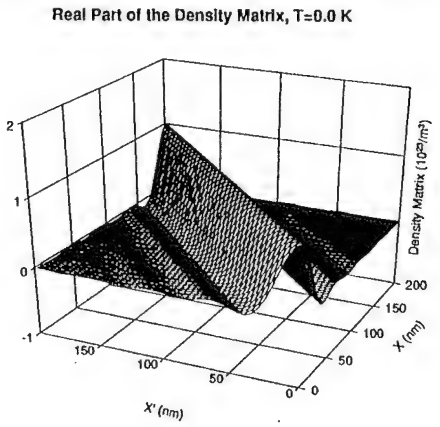


Figure 1. Density matrix for free particles weighted by a Fermi distribution for GaAs at T = 0 K. The density is $10^{23}/m^3$.

The oscillation in the density matrix along the direction (correlation direction) normal to the diagonal is determined by the argument of the spherical Bessel function. The periodicity depends on density as expressed by the Fermi wave number, and suggests the possibility of a wavenumber dependent resonance. The oscillation disappears at room temperature, where the distribution approaches a Gaussian as described by (19). The progressive decrease in the numbers of oscillations as the temperature increases is displayed in Fig. 2, which displays a cut of the density matrix in a plane normal to the diagonal of the density matrix. The effects of Fermi statistics are also more pronounced as the density is increased (e.g., k_F is increased) and we expect this to manifest itself in the oscillatory character of the density matrix.

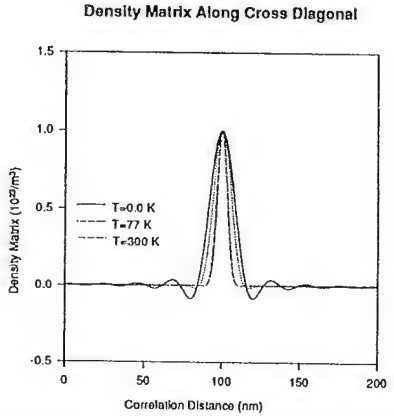


Figure 2. Density matrix versus correlation distance for free particles weighted by a Fermi distribution for GaAs at T = 0 K, 77 K and 300 K.. The density is $10^{23}/m^3$

The density matrix $\rho(x,x')$ shown in Fig. 1 is plotted for a range of values of x and x', $(0 < x < 200nm, 0 < x' < 200nm)$. The density 'observable' $\rho(x) \equiv \rho(x,x)$ is the value of the density matrix along the diagonal and is *plotted* along the physical coordinate x. Pictorially, the density is a *projection* of the diagonal component of the density matrix onto the x-axis. The density matrix along the cross diagonal is defined as $\rho_{cross}(x) \equiv \rho(L-x,x)$, where L is the length of the structure; it is shown as a projection onto the x-axis.

The above discussion provides an indication of what the density matrix coordinate representation profiles corresponding to standard classical equilibrium distribution functions look like. It is expected on physical grounds that a problem examined using the classical distribution function in momentum space would yield the same physical results with respect to the observables in the case of the coordinate representation density matrix. For example, classically, with the Boltzmann distribution, the probability distribution is proportional to $\exp -V(x)/k_B T$. Thus, e.g., when a potential energy equal to $k_B T \ell n 10$ (0.059521 eV at room temperature) is considered, classical theory teaches that the density will be reduced by an order of magnitude from its reference value. Solving the equation of motion of the density matrix for this case provides the same result. If we go to the other extreme at $T = 0$ K, and recognize that the Fermi energy relative to the bottom of the conduction band, $E_F - E_C$, corresponding to a density of $10^{24}/m^3$, is 54.4 mev, while that corresponding to a

density of 10^{23} /m^3, is 11.7 meV, then introducing a barrier of 42.7 meV will reduce the density by an order of magnitude. This is shown in Fig. 3.

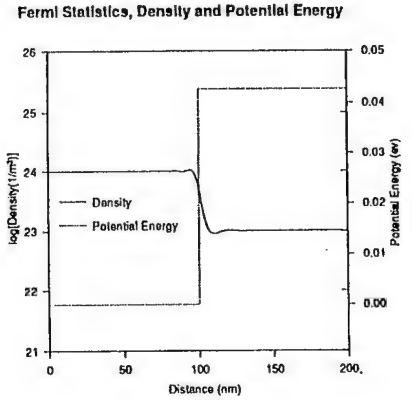


Figure 3. For GaAs at T = 0 K, Fermi statistics, with a step change in potential energy from 0.0 eV to 0.0427 eV (dotted line), the non-self consistent spatial variation in density (solid line).

Apart from the asymptotic (classical) values of density far from the interface, we point to the local oscillation in density on either side of the interface, and make note of the position of the peak and minimum values of density. Classical studies indicate that the peak value of density occurs at the interface, while all quantum mechanical studies indicate that the peak is shifted away from the interface. In a recent density matrix study (Grubin *et al.*, 1993), devoted to Boltzmann statistics, it was analytically demonstrated that the density could be represented in equilibrium as being equal to:

$$\rho(x) = N_c \exp\left[\mu_n - \left(V(x) + Q(x)/3\right)\right] . \qquad (22)$$

In the absence of the quantum potential the density is determined solely by the potential energy, and so the density for the potential energy distribution of Fig. 3 would be equal to one value right up to the potential barrier, and a second (lower) value within the potential barrier. The finite value of the quantum potential and its spatial variation is responsible for the minimum and maximum values of the density occurring away from the interface. This will be discussed in more detail below where we will also illustrate the value of the quantum potential. We will also discuss the factor '3' that appears in (22).

The potential variation in Fig. 3 is imposed and abrupt. Alternatively we can envision a structure in which the density changes abruptly at the same point (100 nm). The solution to the Liouville equation and Poisson's equation yield a potential distribution whose values asymptotically approach those of Fig. 3. The potential distribution at the interface is no longer abrupt, and the local peak seen in Fig. 3 is absent. Rather, there is a more gradual decrease in density across the interface, with values that cannot be described by the classical distribution, but require the presence of the quantum potential. The two dimensional density matrix for the calculations of Fig. 3 are shown in Fig. 4.

The origin of the scales in Fig. 4 is closest to the reader where the density matrix has its highest values. Notice the ripples in the density matrix closest to the highest density regions. Ripples are also present at the lower density regions but their period and magnitude are weaker. Generally the effects of Femi statistics are more pronounced at higher densities, where from (20) it is seen that the amplitude of the oscillation increases, and the period decreases, with increasing density.

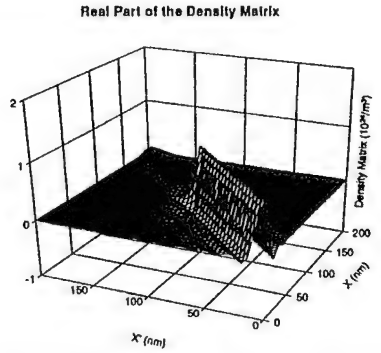


Figure 4. Two dimensional density matrix from which the results of Fig. 3 are obtained.

EQUILIBRIUM DISTRIBUTIONS AND THE QUANTUM POTENTIAL

As indicated in the earlier discussion, the classical distribution function accounts incorrectly for the charge distribution in the vicinity discontinuities in potential energy and cannot be used if the goal is a description of the operational physics of devices. Additionally, we have used the quantum potential as an aid in interpretation. Several cases are treated below which illustrate the significance of the quantum potential. The situation of the resonant tunneling diode will be treated separately where the significance of the quantum potential is most apparent.

The first case of interest is that of a single barrier of modest height, 42.7 mev. This barrier height, we recall, is the value of the step potential of Figs. (3) and (4) where the asymptotic values of density differed by an order of magnitude. For the case illustrated in Fig. 5, we again consider a non self-consistent calculation, with a reference density of 10^{24} /m^3, $T = 0$ K, Fermi statistics and a device length of 200 nm. For the situation when a 42.7 nm barrier, centrally placed and of 100 nm width is considered, it is found that the asymptotic value of density within a central 80 nm region is equal to 10^{23} /m^3, a result expected from the earlier discussion. There was, additionally, the structure in density at the potential discontinuity that was seen in Fig. 3.

When a 10 nm wide barrier is considered, the results are quantitatively different. There is a local peak away from the barrier, but the minimum value of density exceeds that associated with the wider barrier. Of interest, however, is the structure of the quantum potential, shown in Fig. 5. First we note that the magnitudes of $Q(x)$ and $V(x)$ are approxi-

mately the same within the barrier region. The quantum potential is negative within the barrier, a consequence of a positive value of curvature for the density within the barrier (the density reaches a minimum at x = 100 nm). The quantum potential is positive in the regions immediately upstream and downstream of the barrier, where the curvature of the density is negative. The signs of the quantum potential are consistent with a density that is below its classical value immediately outside the barrier, and above its classical value within the barrier region.

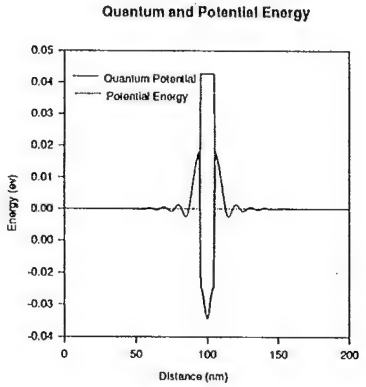


Figure 5. Quantum potential (solid curve) and $V(x)$ (dotted) for a single barrier 10 nm wide.

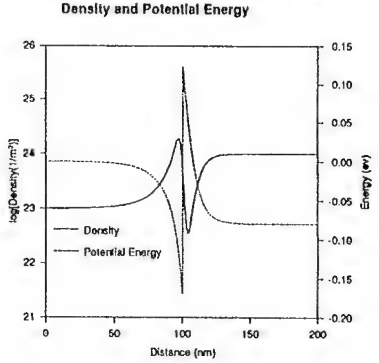


Figure 6. Self-consistent calculation of the density and potential energy for a 300 meV heterostructure diode at T = 300 K, with Fermi statistics and flat band conditions

The next case of interest, which again offers the quantum potential as a vehicle for interpretation, is the familiar self-consistent charge distribution associated with a wide bandgap/narrow bandgap structure. Figures 6 through 8 illustrate results using the density matrix. for a room temperature self-consistent calculation. Here the device length is 200 nm, where for 0<x<100 nm, $N_D = 10^{23}/m^3$, for 100 nm<x<200 nm, $N_D=10^{24}/m^3$. A barrier of 300 mev is imposed. While the non self-consistent calculations of Fig. 3 show a reduction in charge density within and near the edge of the barrier, there is nothing in Fig. 3 resembling the extent of the charge reduction seen in Fig. 6. The contributions to this change are several-fold. *First,* the barrier of Fig. 6 is an order of magnitude higher than that of Fig. 3. *Second,* the applied potential energy difference across the structure is chosen to yield *flat band conditions,* and thus equal to the height of the barrier plus the built-in potential. *Third,* the self-consistent potential displays structure. What is the origin of this structure?

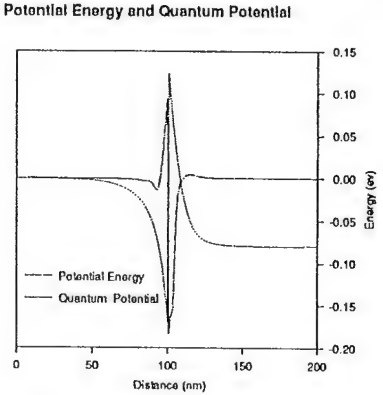


Figure 7. Self-consistent calculation of the quantum potential and potential energy for a 300 meV heterostructure diode at T =300K, with Fermi statistics and flat band conditions

In all of the calculations with a heterostructure barrier, once we pass the peak density there is a progressive decrease in density until a minimum value of density is reached within the interior of the heterobarrier. The simple explanation based upon the quantum potential indicates, from (6) that the net force, under zero current conditions is zero. But *the quantum mechanical self-force, generated by variations in the single particle density (from the quantum potential as seen in Fig. 7) is always nonzero.* Here, as we move into the wide band gap region where the density is decreasing and approaching a minimum value, the curvature of the density is positive, resulting in a negative value for the quantum potential. Since there is a minimum value of the density within the wide band gap region, there is structure to the quantum potential leading to a spatially dependent driving force. This force must be balanced by variations in the self consistent potential, as seen in Fig. 7. The self-consistent potential, which is driven by Poisson's equation, is now subject to the additional constraint imposed by the quantum potential The details are not governed by (6), rather they are governed by the Liouville equation; but the qualitative features are represented by (6). When examining the classical situation we note that the potential energy is also constrained by a *diffusive contribution.* Diffusive contributions are also present when quantum transport is considered. The quantum potential contribution is an additional contribution that is not dependent upon the presence of diffusion.

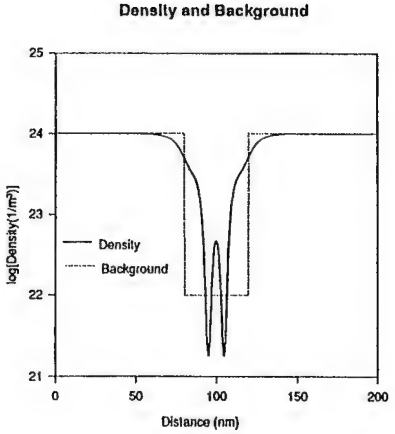


Figure 8. Self-consistent $T = 300$ K calculation with Fermi statistics showing the density and donor distribution for a symmetric double barrier structure.

There are several interesting additional points concerning the structure of the charge distribution associated with the calculations of Figs. 6 and 7. A good approximation to the curvature of the potential energy within the wide band gap region and near the interface, is to assume that the region is free of mobile carriers, whereby $\nabla^2 V(x) = (e^2/\varepsilon)\rho_0(x)$. As a consequence, the higher the heterobarrier, the larger the width of the depletion zone on the wide band gap side of the structure. Under flat band conditions where the net charge distribution is zero, there is a corresponding increase in charge on the narrow band gap side, and this accumulated charge will increase with increasing barrier height. Thus, unlike the non-self consistent calculation of Fig. 3, there is significant charge accumulation on the narrow band side of the structure. The quantum potential, which is negative on the wide band gap side and therefore yields a larger than classical result for the particle density, also has the effect of yielding a lower than classical result for the density just outside of the barrier. The small region of negative quantum potential to the left of the barrier is a consequence of the quantum potential defined in terms of the square root in density. An expansion of the quantum potential leads to contributions from the square of the first derivative of density as well as the second derivative.

What is the situation with multiple barrier structures; the simplest being the double barrier resonant tunneling structure. The characteristic feature of the multiple barrier structures is the existence of quasi-bound states between the barriers. The density between the barriers depends upon the barrier height, barrier configuration, doping, etc. As discussed earlier (Grubin *et al.*, 1994a), the value of the quantum potential within the quantum well of a double barrier structure is approximately equal to the energy of the lowest quasi bound state, relative to the bottom of the conduction band. We note that in terms of the definition of the quantum potential, under steady state, zero current conditions, it is direct to show from Schrodinger's equation that $Q(x)+V(x)=E$, where E is the energy of the quasi-bound state (see also Dirac, 1958). We illustrate the quantum potential for a 200 nm structure double barrier structure. There are two barriers 5nm wide, each 300 meV high, separated by 5 nm, placed in the center of the structure. The background doping is $10^{24}/m^3$ and uniform, except in the interior 40nm region where it is reduced to $10^{22}/m^3$. Figures 8 and 9 show, respectively the density and donor distribution, the quantum potential and the self consistent potential energy.

With respect to Figs. 8 and 9, we note that carriers in excess of $4\times10^{23}/m^3$ reside within the quantum well. The quantum potential is negative within the barriers of the structure corresponding to the curvature of the density, and is positive within the quantum well. But the remarkable feature is that the quantum potential is approximately constant within the quantum well. We have found that for the 300 meV barrier, the quantum potential within the well is approximately 84 meV (for a 200 meV barrier the quantum potential within the well is approximately 70 meV). A key feature in utilizing the density matrix in the coordinate representation is that the quantum potential behaves like a quasi-bound state.

Further evidence for use of the quantum potential within the well as a measure of the energy of the quasi-bound state was provided by supplemental calculations in which the double barrier structure was placed within a 40nm wide quantum well. The depth of the quantum well was varied. As the depth increased the quantum potential between the barriers remained independent of position, but increased slightly in value. The situation when the quantum well was 150mev deep, resulted in a value of the quantum potential between the barriers that increased to 94mev. The detailed results are different than that of figures 8 and 9 in that the density between the barriers has increased (this increase in density has at least two origins: the increased density on either side of the barriers, and the lowering of the quasi-bound state relative to the Fermi level of the entering carriers).

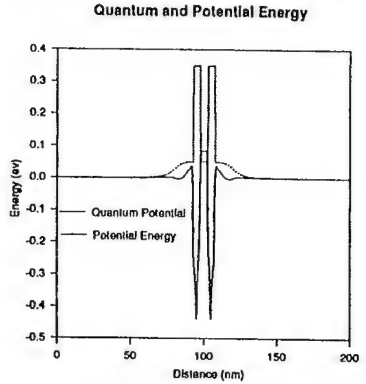


Figure 9. For Fig. 8, the quantum and potential energy distribution.

DISSIPATION AND CALCULATION OF CURRENT

The calculations of the density and potential profiles for the barrier structures in both non self consistent and self consistent studies indicate that for distances sufficiently far from the interface the results are the same as those expected using the dissipationless Boltzmann (or Vaslov) equation. When current flows, classical device transport studies usually proceed

via the drift and diffusion or hydrodynamic equations, or through solutions to the Boltzmann transport equation and Monte Carlo procedures. Here, for cases where the ends of the device are heavily doped N^+ regions, boundary conditions on the numerical procedures are invoked to assure that the numbers of particles leaving and entering the structure are the same. An alternative approach that should yield the same results, with respect to charge and potential energy distributions at the boundaries, is to implement procedures recognizing that dissipation at the beginning and ends of the structure may be represented by carriers that thermalize to a local equilibrium. The issue, then, is how to deal with this situation. To date, very approximate methods have been introduced, and a rational for this approach is discussed below, but it is emphasized that *some procedure for dissipation must be invoked if transport in devices is to be discussed sensibly.*

One of the most succinct ways to express the problem of dissipation follows that of Caldeira and Leggett (1983). We consider a system A (the device), a second system B (the reservoir), and the interaction between the two as described by the Hamiltonian $H_T = H_{Device} + H_{Reservoir} + H_{Interaction}$. The breakup between the device and reservoir is problem dependent. If upper case letters denote the coordinates of the reservoir and lower case letters denote the coordinates of the system of interest (e.g., the electron system) then the quantity interest is the density matrix $<\mathbf{xR}|e^{-iH_T t/\hbar}\rho_{op}(0)e^{-iH_T t/\hbar}|\mathbf{x'R'}>$. This quantity describes the behavior of the entire system. We do not need detailed information about the reservoir; rather, we need to determine its influence on, in our case, the electron system, which implies: $\rho(\mathbf{x},\mathbf{x}',t) = \int d\mathbf{R}d\mathbf{R}' <\mathbf{xR}|e^{-iH_T t/\hbar}\rho_{op}(0)e^{-iH_T t/\hbar}|\mathbf{x'R'}>$.

One method that has been invoked to deal with dissipation and boundaries and current flow in devices, has been guided by perturbation theory on the density matrix (*Grubin et al.*, 1994b). First the equation of motion of the density matrix has been rewritten to include a scattering contribution, as shown by (11). Below we concentrate on the modifications of the Liouville equation through the incorporation of scattering and deal only with *the Liouville equivalent of classical scattering.*

In the Boltzmann picture, ignoring Fermi statistics, the scattering rate is:

$$\left[\frac{\partial f_w(\mathbf{k},\mathbf{x})}{\partial t}\right]_{dis} \equiv -\frac{2}{(8\pi^3)}\int d\mathbf{k}'\{f_w(\mathbf{k},\mathbf{x})W(\mathbf{x},\mathbf{k}',\mathbf{k}) - f_w(\mathbf{k}',\mathbf{x})W(\mathbf{x},\mathbf{k},\mathbf{k}')\} \ , \tag{21}$$

where the subscript 'w' denotes a Wigner function and $W(\mathbf{x},\mathbf{k},\mathbf{k}')$ represents the standard transition probability per unit time. Utilizing the Weyl transformation:

$$\rho(\mathbf{r}+\mathbf{s},\mathbf{r}-\mathbf{s}) = \frac{2}{8\pi^3}\int d\mathbf{k}\, f_w(\mathbf{k},\mathbf{r})\exp i2\mathbf{s}\cdot\mathbf{k} \ , \tag{22}$$

with the following change in coordinates: $\mathbf{x}+\mathbf{x}' = 2\mathbf{r}$, $\mathbf{x}-\mathbf{x}' = 2\mathbf{s}$, the scattering rate of density matrix (after manipulation of the variables of integration) is given by:

$$\left[\frac{\partial \rho(\mathbf{r}+\mathbf{s},\mathbf{r}-\mathbf{s})}{\partial t}\right]_{scat} \equiv -\left[\frac{2}{8\pi^3}\right]^2\int d\mathbf{k}d\mathbf{k}'\{f_w(\mathbf{k},\mathbf{r})[\exp i2\mathbf{s}\cdot\mathbf{k}]W(\mathbf{r},\mathbf{k}',\mathbf{k})\}\{1-\exp i2\mathbf{s}\cdot(\mathbf{k}'-\mathbf{k})\} \tag{23}$$

The structure of the scattering term within the framework of classical Boltzmann scattering expressed within the coordinate representation is obtained from (23). For example, the second exponential term in (23) can be expressed as an infinite series, in which case the scattering term is expressed as an infinite series in powers of s. The lead term is given by:

$$\left[\frac{\partial \rho(\mathbf{r}+\mathbf{s},\mathbf{r}-\mathbf{s})}{\partial t}\right]_{scat} \equiv -i2\mathbf{s}\cdot\left[\left[\frac{2}{8\pi^3}\right]^2\int d\mathbf{k}\, f_w(\mathbf{k},\mathbf{r})[\exp i2\mathbf{s}\cdot\mathbf{k}]\int d\mathbf{k}'(\mathbf{k}'-\mathbf{k})W(\mathbf{r},\mathbf{k}',\mathbf{k})\right] \tag{24}$$

Standard classical theory (Ferry, 1991) teaches that:

$$\frac{2}{(8\pi^3)}\int d\mathbf{k}'(\mathbf{k}'-\mathbf{k})W(\mathbf{r},\mathbf{k}',\mathbf{k}) \equiv -\mathbf{k}\Gamma(\mathbf{r},|\mathbf{k}|) \ , \tag{25}$$

where $\Gamma(\mathbf{r},|\mathbf{k}|)$ represents a scattering rate. Thus:

$$\left[\frac{\partial \rho(\mathbf{r}+\mathbf{s},\mathbf{r}-\mathbf{s})}{\partial t}\right]_{scat} \equiv -i2\mathbf{s}\cdot\left[\frac{2}{8\pi^3}\int d\mathbf{k}\, f_w(\mathbf{k},\mathbf{r})[\exp i2\mathbf{s}\cdot\mathbf{k}]\mathbf{k}\Gamma(\mathbf{r},|\mathbf{k}|)\right] \ , \tag{26}$$

which, using the inverse of the Weyl transformation:

$$2f_w(\mathbf{k},\mathbf{r}) = 2^3\int d\mathbf{s}\ \rho(\mathbf{r}+\mathbf{s},\mathbf{r}-\mathbf{s})\exp -i2\mathbf{s}\cdot\mathbf{k} \tag{27}$$

can be rearranged as

$$\left[\frac{\partial \rho(\mathbf{r}+\mathbf{s},\mathbf{r}-\mathbf{s})}{\partial t}\right]_{scat} \equiv -i2\mathbf{s}\cdot\left[\frac{2^3}{8\pi^3}\int d\mathbf{k}\, d\mathbf{s}'\ \rho(\mathbf{r}+\mathbf{s}',\mathbf{r}-\mathbf{s}')[\exp i2(\mathbf{s}-\mathbf{s}')\cdot\mathbf{k}]\mathbf{k}\Gamma(\mathbf{r},|\mathbf{k}|)\right] \tag{28}$$

A significant simplification arises when the crystal momentum in (28) is replaced by a divergence of the correlation vector:

$$\left[\frac{\partial \rho(\mathbf{r}+\mathbf{s},\mathbf{r}-\mathbf{s})}{\partial t}\right]_{scat} \equiv -i\nabla_{\mathbf{r}}\cdot\left[\frac{2^3}{8\pi^3}\int d\mathbf{k}\, d\mathbf{s}'\ \rho(\mathbf{r}+\mathbf{s}',\mathbf{r}-\mathbf{s}')[\exp i2(\mathbf{s}-\mathbf{s}')\cdot\mathbf{k}]\Gamma(\mathbf{r},|\mathbf{k}|)\right] \tag{29}$$

For the case when the scattering rate is independent of momentum, the dissipation term reduces to:

$$\left[\frac{\partial\,\rho(\mathbf{r}+\mathbf{s},\mathbf{r}-\mathbf{s})}{\partial\,t}\right]_{scat} \equiv -\Gamma\mathbf{s}\cdot\nabla_{\mathbf{r}}\,\rho(\mathbf{r}+\mathbf{s},\mathbf{r}-\mathbf{s}) \ , \tag{30}$$

and the Liouville equation in the coordinate representation is modified to read:

$$\begin{aligned} i\hbar\frac{\partial\rho(\mathbf{x},\mathbf{x}',t)}{\partial\,t} &= \\ -\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial\,\mathbf{x}^2}-\frac{\partial^2}{\partial\,\mathbf{x}'^2}\right)\rho(\mathbf{x},\mathbf{x}',t)+\left(V(\mathbf{x})-V(\mathbf{x}')\right)\rho(\mathbf{x},\mathbf{x}',t) \ . \\ -i\hbar\frac{\Gamma}{2}(\mathbf{x}-\mathbf{x}')\cdot(\nabla_{\mathbf{x}}-\nabla_{\mathbf{x}'})\,\rho(\mathbf{x},\mathbf{x}') \end{aligned} \tag{31}$$

The additional contribution due to dissipation was discussed in Grubin *et al.* (1993) and in a study by Dekker (1977). Density matrix algorithms incorporating the dissipation contributions of (31) have been implemented with some results reported (Ferry, 1991). But, because of numerical difficulties at higher bias levels, modifications to the scattering were introduced whose consequences go beyond the approximations associated with the expansion of (24). It is worthwhile dwelling on these modifications.

In modifying the scattering term in (31) it was recognized that the dissipation term could be re-expressed in terms of a velocity density matrix:

$$\mathbf{j}(\mathbf{x},\mathbf{x}')=\left(\frac{\hbar}{2mi}\right)(\nabla_{\mathbf{x}}-\nabla_{\mathbf{x}'})\,\rho(\mathbf{x},\mathbf{x}') \ . \tag{32}$$

The diagonal elements of (32) yield the velocity flux density. In terms of $\mathbf{j}(\mathbf{x},\mathbf{x}')$, (30) becomes:

$$\begin{aligned} i\hbar\frac{\partial\rho(\mathbf{x},\mathbf{x}',t)}{\partial\,t} &= \\ -\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial\,\mathbf{x}^2}-\frac{\partial^2}{\partial\,\mathbf{x}'^2}\right)\rho(\mathbf{x},\mathbf{x}',t)+\left(V(\mathbf{x})-V(\mathbf{x}')\right)\rho(\mathbf{x},\mathbf{x}',t)+m\Gamma(\mathbf{x}-\mathbf{x}')\cdot\mathbf{j}(\mathbf{x},\mathbf{x}') \end{aligned} \tag{33}$$

The scattering term in the above equation was then written in the form of a scattering potential. The procedures for this were as follow. *First,* the term $\mathbf{j}(\mathbf{x},\mathbf{x}')$ was rewritten as $\mathbf{j}(\mathbf{x},\mathbf{x}')\equiv\mathbf{v}(\mathbf{x},\mathbf{x}')\,\rho(\mathbf{x},\mathbf{x}')$, where $\mathbf{v}(\mathbf{x})\equiv\mathbf{v}(\mathbf{x},\mathbf{x})$ represents the expectation value of the velocity. *Second,* $\mathbf{j}(\mathbf{x},\mathbf{x}')$ was approximated as $\mathbf{j}(\mathbf{x},\mathbf{x}')\approx\mathbf{v}(\mathbf{x})\,\rho(\mathbf{x},\mathbf{x}')$. Higher order terms are at least second order in (x-x'), and retaining them would be inconsistent with the approximation leading to (24). *Third,* quasi-Fermi levels were introduced through the definition:

$$E_F(\mathbf{x})-E_F(\mathbf{x}')=-\int_{\mathbf{x}'}^{\mathbf{x}}d\mathbf{x}''\cdot\mathbf{v}(\mathbf{x}'')m\Gamma(\mathbf{x}'') . \tag{34}$$

For small values of x-x' about x, (34) is approximately represented by: $E_F(\mathbf{x})-E_F(\mathbf{x}')=-(\mathbf{x}-\mathbf{x}')\cdot\mathbf{v}(\mathbf{x})m\Gamma(\mathbf{x})$. Under this approximation, (33) becomes

$$\begin{aligned} i\hbar\frac{\partial\rho(\mathbf{x},\mathbf{x}',t)}{\partial\,t} &= \\ -\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial\,\mathbf{x}^2}-\frac{\partial^2}{\partial\,\mathbf{x}'^2}\right)\rho(\mathbf{x},\mathbf{x}',t)+\left[\left(V(\mathbf{x})-V(\mathbf{x}')\right)-\left(E_F(\mathbf{x})-E_F(\mathbf{x}')\right)\right]\rho(\mathbf{x},\mathbf{x}',t) \end{aligned} \tag{35}$$

Thus, we have taken the differential (31) whose right hand side is complex, and replaced it by one whose right hand side is real, when the density matrix is real. Side by side calculations at low values of bias yield identical results. While the above discussion leading to (35) appears to be model dependent, the results implied by this equation have greater generality than the means used to arrive at it.

The implementation of (35) permits us to calculate current in a direct manner. How is this done? In all of the calculations with current the assumption is that the carriers at the upstream boundary are in local equilibrium and that the distributions are either a displaced Maxwellian or a displaced Fermi-Dirac distribution. As discussed in Grubin *et al.* (1993), this implies that at the upstream boundary the zero current quantum distribution function $\rho(\mathbf{x},\mathbf{x}')$ is replaced by $\rho(\mathbf{x},\mathbf{x}')\exp[im\mathbf{v}(boundary)\cdot(\mathbf{x}-\mathbf{x}')/\hbar]$. *Since current is introduced as a boundary condition to the problem, as formulated by (35), a prescription is necessary for finding its value. An auxiliary condition was constructed.*

To compute a value of current for use in the Liouville equation, a *criteria* was introduced through moments of (35) (Grubin *et al.*, 1993). Under time independent conditions, the momentum balance equation yields the condition

$$2\nabla_{\mathbf{x}}E+[\nabla_{\mathbf{x}}V]\rho(\mathbf{x})-[\nabla_{\mathbf{x}}E_F]\rho(\mathbf{x})=0 \ , \tag{36}$$

where E is the kinetic energy and is given by (13). Under the assumptions of current continuity, i.e., $\rho(\mathbf{x})\mathbf{v}(\mathbf{x})$ is independent of distance (satisfied for the Liouville equation), *and the condition that the energy of the entering and exiting carriers are* equal, (34) becomes

$$E_F(x)-E_F(x')=-j\int_{x'}^{x}dx''m\Gamma(x'')/\rho(x'') , \tag{37}$$

where we have restricted the considerations to one space dimension. *The current is chosen so that $E_F(L)-E_F(0)$ is equal to the change in applied energy across the structure.*

We now illustrate some of the above considerations. The simplest type of calculation to deal with is that of a free particle. For this case, and with current introduced as a boundary condition, the density matrix is complex. The real part is symmetric and the imaginary part (from which current is obtained) is asymmetric about the diagonal. The calculation displayed in Fig. 10 shows the real part and Fig. 11 the imaginary part for a 200nm with a doping of $10^{23}/m^3$ subject to a bias of 10 meV. For this calculation and parameters appropriate to GaAs, a scattering rate of 10^{13} sec yields a mobility of 0.258 m^2/V-sec. The mean carrier velocity for this calculation is approximately 1.3×10^4 m/sec.

Increasing the applied bias results in an increase in the carrier velocity and an increase in the kinetic energy of the carriers. This increase affects the curvature of the density matrix in the correlation direction and is displayed in Fig. 12.

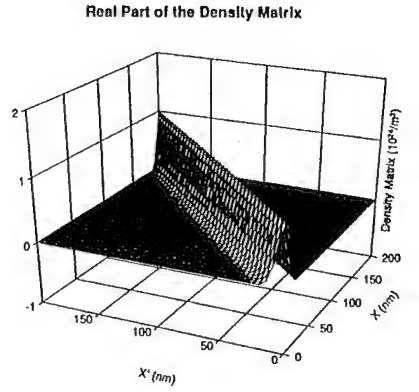


Figure 10. Real part of the density matrix for a free particle subject to a constant force.

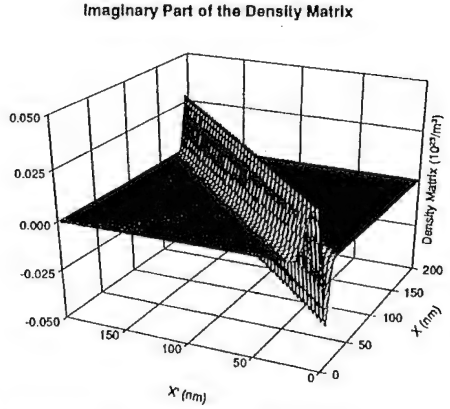


Figure 11. Imaginary part of the density matrix corresponding to Fig. 10.

All semiconductor devices sustain energy dependent scattering, implying that the scattering rate within one region of the structure will be different than at a different region of the structure. To understand how this is implemented in the density matrix algorithm, several illustrative examples of nonuniform scattering were performed. These examples deal with the generation of nonuniform fields from variation in the *mobility* (vis scattering). We will treat an element with material parameters nominally the same as those associated with Figs. 10-12. However, here we vary the scattering rate within the central 5 nm of the structure. On the basis of the definition of the quasi-Fermi energy, a decrease in the scattering time, which results in a decrease in mobility, will yield a sharp drop in the quasi-Fermi level. The density cannot change as rapidly, but is constrained by the Debye length, and results in a more gradual change in the self consistent potential energy. The quasi-Fermi energy and potential energy, as well as the density, are displayed in Fig. 13 for a bias of 10 meV, where the scattering time within the central 5 nm was 10^{-14} sec, while that at the boundaries is 10^{-12} sec. There are several points to emphasize. For the calculation of Fig. 13, the quasi-Fermi energy varies in an approximately linear manner in three separate regions. In particular, within the exterior cladding regions the quasi-Fermi level is equal to the potential energy distribution, where it assures the presence of local charge neutrality. The departure of the potential energy from the quasi-Fermi energy for this calculation is in large part a consequence of Debye length considerations. The quasi-Fermi energy, which is an integral expression follows the same slope to the interior region, where the precipitous change in value is a consequence of the reduction in the scattering time.

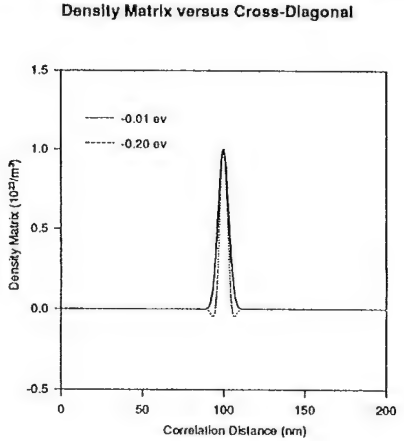


Figure 12. Density matrix versus correlation distance when current is flowing. Dashed line is for a bias of 10 meV and a mean velocity of 1.3×10^4 m/sec; solid line is for a bias 200 meV and a mean velocity of 2.6×10^5 m/sec.

Figure 14 displays the scattering rate used in the calculations and the self-consistent density distribution. Of extreme significance here is the formation of a local dipole layer within the interior of the structure.

SINGLE BARRIER DIODE: CONSTANT SCATTERING RATE

The quasi-Fermi scattering model has been applied to a variety of structures, including single and multiple barrier diodes, as well as electron-hole transport. We illustrate single barrier calculations in Figs. 15 through 18 for a structure with a constant scattering rate. Preliminary results for this type of structure were presented earlier (Ferry and Grubin, 1993). The calculations are for a 200 nm structure containing a single 300 meV high, 20 nm wide barrier embedded within a 30nm N^- region, surrounded by uniformly doped $10^{24}/m^3$ material. The scattering time, τ, is constant and equal to 10^{-13} sec. The calculations are self-consistent and assume Fermi-Dirac boundary conditions. The first three figures, 15 through

17, show potential energy, density, and quasi-Fermi energy distributions, respectively, for different bias levels.

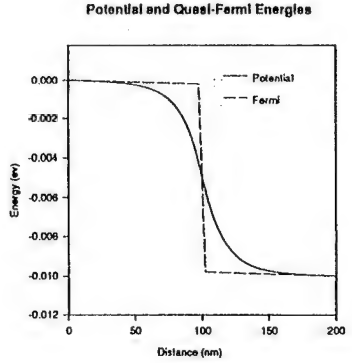


Figure 13. Self-consistent calculations of the potential energy and quasi-Fermi energy for a uniform doped structure with a variable scattering rate within the center of the structure.

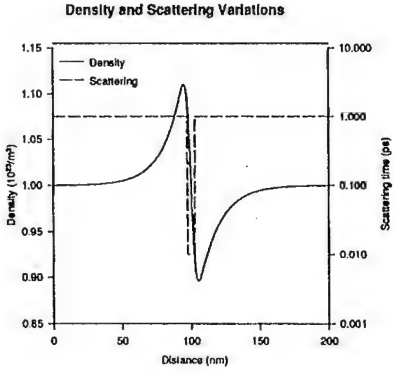


Figure 14. Self-consistent calculation of the density for a uniform doped structure with the displayed variable scattering rate.

From Fig. 15, as the collector boundary is made more negative with respect to the emitter, a local 'notch' potential well forms on the emitter side of the barrier. The potential energy decreases linearly across the barrier, signifying negligible charge within the barrier, followed by a broad region where the potential energy decreases to its value at the collector boundary.

The charge distribution, Fig. 16, displays a buildup of charge on the emitter side of the barrier, and a compensatory region of charge depletion on the collector side of the barrier. At a bias of 400 mev significant charge accumulation has formed on the emitter side of the barrier, followed by a broad region of charge depletion on the collector side. Note that as the bias increases there is a progressive increase in charge within the interior of the barrier. Both results are consistent with the low temperature experimental findings of *Eaves et al.* (1990).

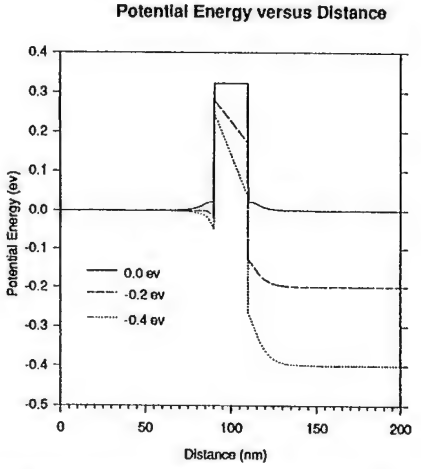


Figure 15. Self consistent room temperature potential energy calculations assuming Fermi-Dirac boundary conditions for a single barrier structure under varying bias conditions.

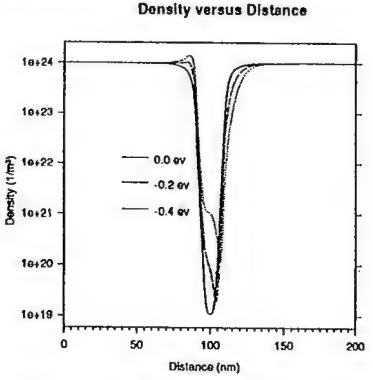


Figure 16. Self consistent density calculations for Fig. 15.

Figure 17 displays the quasi-Fermi energy (relative to the equilibrium Fermi energy). Because of the low values of current E_F, is approximately zero from the emitter to within the first half of the barrier (in the emitter, the variation in the Fermi level matches that of the

potential and insures that the density matrix is constant near the boundary) and then drops to a value approximately equal to the bias through the remaining part of the structure.

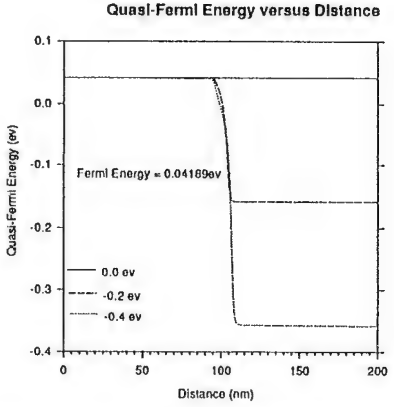


Figure 17 Quasi-Fermi energy distribution for the calculations of Fig. 15.

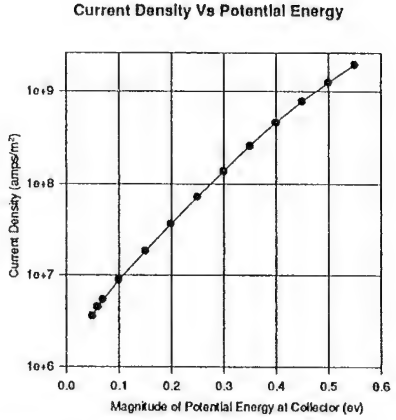


Figure 18. Current-voltage relation for the calculations of Figs. 15 through 17.

I versus V for the 20 nm barrier is shown in Fig. 18. Note that for a broad range of voltage the current depends exponentially on voltage; but there is distinct sublinearity to the curve. In other words, the sublinearity indicates that at a given value of voltage the current is lower than expected on the basis of a pure exponential relation. In seeking an origin of this sublinearity, we note from the accompanying voltage distributions that not all of the voltage falls across the tunnel barrier; indeed, a substantial contribution falls across the region immediately adjacent to the collector side of the barrier.

RESONANT TUNNEL DIODE; VARIABLE SCATTERING RATE.

To illustrate the calculation for resonant tunneling structures, we treat a 200 nm structure, with two 5 nm, 300 meV barriers separated by a 5 nm well. The doping is 10^{24} /m^3 except for a central 50 nm wide region where the doping is 10^{22} /m^3. The effective mass is constant and equal to that of GaAs ($0.067m_0$); Fermi statistics are used; the ambient is 77 K; and current is imposed through the density matrix equivalent of a displaced distribution at the boundaries. In these computations one set of scattering rates was used, *although scattering was increased in the vicinity of the double barriers.*

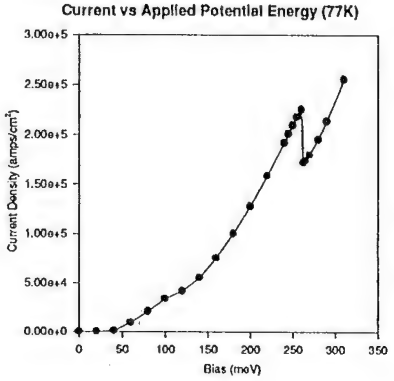


Figure 19. Current versus (magnitude) voltage for the resonant tunneling structure.

The signature of the RTD is its current-voltage relation with the region of negative differential conductivity; this is displayed in Fig. 19. The current is numerically negligible until a bias of approximately 50 meV, with the peak current occurring at 260 meV, followed by a sharp but modest drop in current at 270 meV. The interpretation of these results is assisted by Figs. 20 and 21 and the Bohm quantum potential. As indicated earlier, we found, through an extensive number of numerical simulations, that the value of $V(x)+Q(x)$, between the barriers of an RTD is a measure of the position of the quasi-bound state.

Figure 20 displays the equilibrium self-consistent potential for the RTD. Also shown is the equilibrium Fermi energy (approximately 54 mev) and five different values of applied potential energy, of $V(x)+Q(x)$ within the quantum well. At 100 meV the quasi-bound state is approximately equal to the equilibrium Fermi energy and significant current begins to flow. The current increases until a bias of 260 meV, where there is a current drop.

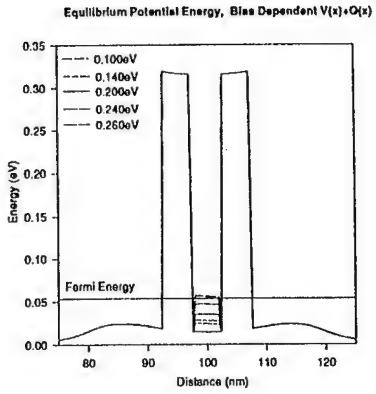


Figure 20. Equilibrium potential energy and the bias dependence of $V(x)+Q(x)$ within the quantum well. Legend denotes collector bias.

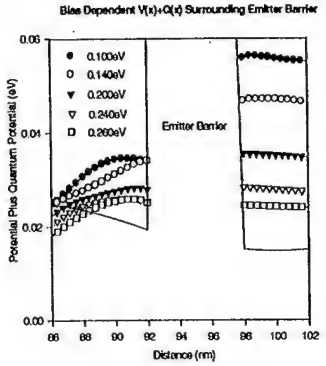


Figure 21. Blow up of Fig. 20.

To see what is happening, we blow up the region on either side of the emitter barrier, where we display values of $V(x)+Q(x)$ before the emitter barrier and within the quantum well (Fig. 21). Within the quantum well we see the quasi bound state decreasing as the bias on the collector is increasing. In the region prior to the emitter barrier, where a 'notch' potential forms signifying charge accumulation, we see the formation with increased bias of a region where $V(x)+Q(x)$ is relatively flat. Of significance here is that for values of bias associated with the initial current increase the value of $V(x)+Q(x)$ within the quantum well is greater then its value in the emitter region. The current reaches a maximum at the crossover, where $V(x)+Q(x)$ in the emitter region and in the quantum well are approximately equal. (Implementation of an earlier algorithm, generally resulted in solutions oscillating between high and low values of current when this condition was reached). While it is tempting to associate $V(x)+Q(x)$ within the emitter region with a quasi-bound state, this may be premature.

The distribution of potential energy, $V(x)$, as a function of bias is displayed in Fig. 22, where the notch potential is deepened with increasing bias, signifying increased charge accumulation. This is accompanied by a smaller share of the potential drop across the emitter barrier, relative to the collector barrier region. Comparing the slopes of the voltage drop across the emitter and collector barriers, we see larger fractions of potential energy fall across the collector barrier.

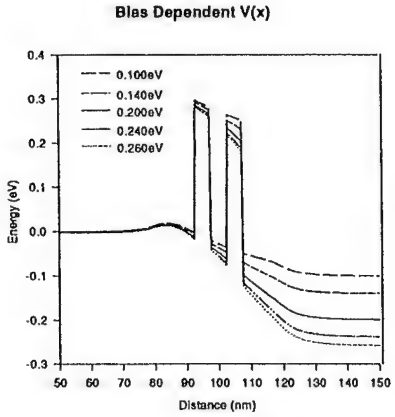


Figure 22. Distribution of potential energy as a function of applied bias.

Explicit in this calculation is dissipation, which is incorporated through the quasi-Fermi level. Within the vicinity of the boundaries the quasi-Fermi level is parallel to the conduction band edge. Indeed, for this calculation the quasi-Fermi level departs from the conduction band edge only within the vicinity of the barriers. The quasi-Fermi level is displayed in Fig. 23 at a bias of 260 meV, where we see that the quasi-Fermi level is relatively flat until the middle of the first barrier at which point there is a small drop in value, followed by a flat region within the quantum well. There is a strong drop of the quasi-Fermi level within the second barrier.

The charge distribution accompanying these variations in bias shows accumulation on the emitter side of the barrier, along with charge accumulation within the quantum well. The increase in charge within the quantum well and adjacent to the emitter region is accompanied by charge depletion downstream of the second barrier, with the result that the net charge distribution throughout the structure is zero.

Variations in the quasi Fermi level were accompanied by variations in density and current, which were all obtained in a self-consistent manner. Supplemental computations were performed, in which the quasi-Fermi level was varied by altering the scattering rates. The calculations were applied to the post threshold case, with values for the scattering rate

chosen so to provide a large drop in current. Indeed, a current drop by greater than a factor of three was obtained, followed by a shallow current increase with increasing bias. The significant difference leading to these changes was the manner in which the quasi-Fermi level changed. Rather than the shallow change depicted in Fig. 19, there was a larger change in the quasi-Fermi level across the first barrier, a result similar to that obtained for single barriers.

The calculations obtained for Figs. 19 through 23 were obtained from a new solution algorithm that was constructed for the quantum Liouville equation that permits a more convenient specification of boundary conditions, in particular, when the device is under bias. The algorithm is based on a reformulation of the governing equations, in which a higher order differential equation in the local direction $[(x+x')/2]$ is constructed from the quantum Liouville equation. The reformulated equation behaves like an elliptical equation in the local direction, rather than the hyperbolic behavior of the quantum Liouville equation. With appropriate boundary conditions, solutions to the two forms of the quantum Liouville equations are equivalent. However, the reformulated equation allows construction of a more robust algorithm that provides desired solution behavior at the contacts by boundary condition specification at both contacts.

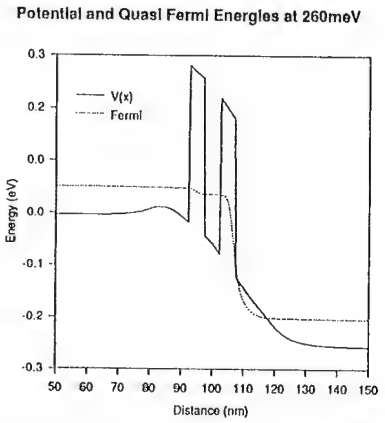


Figure 23 Potential and quasi-Fermi energy at a bias of 260 meV.

DIFFERENTIAL CAPACITANCE VERSUS VOLTAGE CALCULATIONS

Differential capacitance versus voltage (CV) measurements permit a reconstruction of density profiles in semiconductor heterostructures based upon a standard textbook formula: $N_{reconstructed}(< x >) \propto [dC^{-2}/dV]^{-1}$, where $< x >= \varepsilon / C$. Further, the extrapolated intercept of $1/C^2$ versus bias V yields the offset voltage for simple heterostructure configurations. Thus, from a single set of measurements qualitative information regarding the position of the key heterointerfaces, and quantitative information concerning the offset voltages are obtained.

For the past decade CV studies have also included a numerical component involving the solution of Poisson's equation for a density distribution computed for a specific heterostructure configuration, from which *computed CV characteristics are obtained.* The theoretical structure providing the closest fit between the calculated and measured CV relation and consequent reconstructed density is often pronounced as the one representing the actual heterostructure. The degree to which theory is a reliable guide to CV measurements is dependent upon: (a) the fundamental equations chosen to represent the structure (quantum structures require equations describing quantum transport); (b) the statistics, either Boltzmann or Fermi statistics (first order contributions influence the intersection of the capacitance curve and the bias axis, providing critical information about the offset voltages.); (c) traps and unusual doping contributions (e.g. planar doping); (d) specific quantum boundary effects. In short the most representative calculation is that with the most physics. Thus, the equations least likely to offer confidence in the reconstructed calculations are based upon classical equations, e.g., the drift and diffusion equations. The equations most likely to offer confidence are those yielding quantum distribution functions, such as the Wigner function or the density matrix. The capacitance is obtained via Kroemer *et al.* (1980), as follows.

From Poisson's equation the change in density, $\delta\rho(x)$, subject to the change in applied potential energy, $\delta V(L)$, at $x=L$, yields a net change in charge density:

$$\delta Q(L) = -e\int_0^L dx\,\delta\rho(x). \tag{39}$$

The differential capacitance is:

$$C(V) = \frac{\varepsilon\delta Q(L)}{\delta V(L)} \equiv \frac{\varepsilon}{< x >}, \tag{40}$$

which provides an implicit definition of <x>. From Poisson's equation:

$$< x >= \frac{\delta V}{\delta\left[dV(x=L)/dx\right]}. \tag{41}$$

and the 'width' of the capacitor is determined by the ratio of the change in applied potential energy at the boundary to the change in field at the boundary.

The carrier density is reconstructed from standard textbook expressions (Schroder, 1990)

$$N_{reconstructed} = 2N_{ref}L_D^2\left(\frac{d < x >^2}{d\beta V}\right)^{-1}, \tag{42}$$

where $\beta = (k_B T)^{-1}$, and L_D is the Debye length

$$L_D^2 \equiv \frac{\varepsilon}{\beta e^2 N_{ref}},$$

and N_{ref} is a reference density. For a uniform structure $N_{reconstructed} = N_{ref}$.

Equation (42) is routinely used as an adjunct to experiments, to obtain information about doping profiles, offset voltages, etc. A reverse analysis is also pursued where an assumed device configuration is assumed and the *C-V* relationship is obtained. The resulting *C-V* is then compared to experiment where the closest fit is pronounced as the design of the structure under study. Thus, as indicated above, the numerical results are dependent upon the physics used to represent the device under consideration.

In the calculation discussed below the Liouville equation in the coordinate representation was coupled to Poisson's equation For zero current *CV* was obtained for a 200 nm long structure, nominally doped to $10^{24}/m^3$, with a 15 nm, 300 meV barrier within a 30 nm, N^- region. The density and potential energy at equilibrium are shown at equilibrium Fig. 24. The reconstructed density and the density computed from the quantum Liouville equation in the coordinate representation are shown in Fig. 25.

There are several points of note. First, the minimum value of the reconstructed is orders of magnitude higher than that computed from the density matrix, although on a linear plot the apparent difference would appear to be smaller. Second there is a region in which the minimum in the reconstructed density is smaller than that obtained from the Liouville equation. Again, while the net charge density has not be computed it appears that the integrated charge density obtained from the density matrix, and that from the reconstructed density are the same.

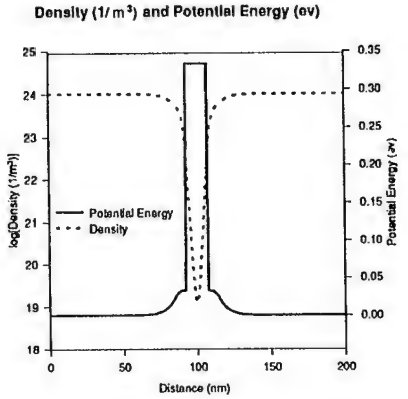


Figure 24. Equilibrium distribution of density from the Liouville equation, and potential energy for a single barrier located within an N^- region.

The reconstructed density is obtained from the derivative of $\langle x \rangle^2$ versus applied bias, which is shown in Fig. 26. According to (42), the slope of $\langle x \rangle^2$ determines the reconstructed density. In the calculation, the central region was approximately three orders of magnitude smaller in doping than the cladding region. Thus if the low doped region was long enough for the density to approach its background value, the slope of $\langle x \rangle^2$ versus $V_{applied}$ would be correspondingly three orders of magnitude higher at the higher bias levels than at the low bias levels. Debye tailing would determine that rate at which this high slope would be reached. For the situation of Figs. 24 through 26, with particular attention to Fig. 26, where the normalized $\langle x \rangle^2$ versus $V_{Applied}$ is displayed, the rate at which the slope is increased is accelerated by the presence of the barrier. And it is anticipated that increases in the barrier height would result in a larger slope. In Fig. 26, we have drawn a line tangent to the slope within the depleted region. The intersection of this line with the two dotted lines provides a measure of the bias need to move $\langle x \rangle$ across the barrier-plus-N^- region. Thus it appears that the CV measurement will yield information about the position of the barrier; it is not clear that information about the height of the barrier can be obtained using this technique.

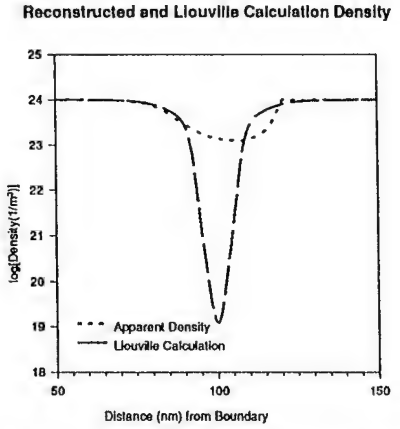


Figure 25. Reconstructed density and density from the Liouville equation for a single barrier structure with a 30 nm N^- region in the center of the structure.

THE QUANTUM HYDRODYNAMIC EQUATIONS

In the discussion of dissipation, several steps were considered. In one case the consequences of scattering were initiated through a transformation of scattering within the framework of the Boltzmann transport equation. Because of limitations of the approximate model, we found it necessary to introduce an auxiliary condition to assure that the carriers at the entering and exiting contacts satisfied dynamic flat band conditions. The condition imposed was that the kinetic energy density and density of the entering and exiting carriers were equal. The implication is that the entering and exiting carriers possess the same mean velocity, and have thermalized to a local equilibrium temperature. The specific conditions imposed on the quasi-Fermi energy were obtained from the quantum hydrodynamic equations.

The quantum hydrodynamic equations which were introduced several years ago within the context of device (Ancona and Iafrate, 1989; Grubin and Kreskovsky, 1989) have taken on a life of their own, with a special issue including a series of key papers (there will be a special issue of the *IEEE Trans. VLSI*). These equations also provide some insight into dealing with the Liouville equation in more than one direction. The standard means of dealing with hydrodynamic equations, particularly those obtained from either the Boltzmann or Wigner equations, is to multiply either of these equations by a power of momentum and

integrate over all momentum states. The zeroth moment, first moment and second moment equations yield, respectively the continuity, momentum and energy balance equations. The density matrix in the coordinate representation, which is related to the Wigner function through an integral transformation, is a solution to the Liouville equation in the coordinate representation. *As a consequence, the quantum hydrodynamic equations are obtained from this Liouville equation through a series of derivatives in the correlation direction.* As discussed above, the first correlation derivative yields velocity flux density, while the second correlation derivative yields the energy density.

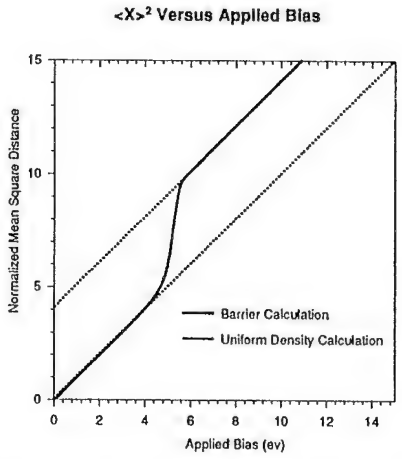


Figure 26. Normalized mean square distance versus voltage for the calculation of Fig. 24.

Since all balance equations discussed in the past have included dissipation when examined within the context of device physics, it is essential that the quantum Liouville equation used for obtaining the balance equation contain dissipation. In the previous section we discussed dissipation and its model dependent nature. Below we consider two such models for the quantum Liouville equation and its consequences for the moment equations.

For dissipation in the form of simple relaxation time approximation, the density matrix equation is

$$\frac{\partial \rho}{\partial t}+\left[\frac{\rho-\rho_o}{\tau}\right]+\left[\frac{\hbar}{2mi}\right]\left[\nabla_r\bullet\nabla_s\right]\rho+\left(\frac{i}{\hbar}\right)V(\mathbf{r}+\mathbf{s})-V(\mathbf{r}-\mathbf{s})]\rho=0\ , \tag{43}$$

whereas for Fokker-Planck dissipation the relevant equation is:

$$\frac{\partial \rho}{\partial t}+\left[2\gamma\mathbf{s}\bullet\nabla_s\rho+\left(\frac{4D\mathbf{s}\cdot\mathbf{s}}{\hbar^2}\right)\rho\right]+\left[\frac{\hbar}{2mi}\right]\left[\nabla_r\bullet\nabla_s\right]\rho+\left(\frac{i}{\hbar}\right)V(\mathbf{r}+\mathbf{s})-V(\mathbf{r}-\mathbf{s})]\rho=0\ . \tag{44}$$

The zeroth moment equation, or the continuity equation, is obtained by taking the diagonal limit of (43) and/or (44). For (44), this yields, using the definition of velocity flux density in one dimension, (14):

$$\frac{\partial\ \rho(x)}{\partial t}+\frac{\partial\ j(x)}{\partial\ x}=0\ , \tag{45}$$

where we have replaced the velocity flux, $j(x)$, by $\rho(x)v(x)$

The first moment equation is obtained by taking the first derivative of the quantum Liouville equation. In the diagonal limit the first moment equation using the definition of energy density, (15) yields for Fokker-Planck dissipation,

$$\frac{\partial\ \rho(x)mv(x)}{\partial t}+2\gamma\ \rho(x)mv(x)+\frac{\partial\ 2E(x)}{\partial x}+\rho(x)\frac{\partial V(x)}{\partial x}=0\ . \tag{46}$$

Equation (46) is the classical form of the first moment equation. All quantum corrections come in through the form of the energy density. Classically, and for Boltzmann statistics the energy density is given by $\rho(x)k_BT/2$. For a constant force, where $\partial V(x)/\partial x$ is constant, and for a spatially constant density, equation (46) reduces to Newton's equation of motion with frictional damping.

In introducing quantum corrections one direct approach is to follow Wigner (932) and introduce the Wigner potential energy:

$$Q_w(x)\equiv-\frac{\hbar^2}{8m}\frac{\partial^{\ 2}\ell n\rho(x)}{\partial\ x^2}\ . \tag{47}$$

Small approximations have shown that the Wigner potential energy is a quantum correction to the classical energy density. Indeed, for the energy associated with one degree of freedom

$$E(x)=\left[\frac{mv^2}{2}+\frac{k_BT}{2}+\frac{Q_w}{3}\right]\rho(x)\ . \tag{48}$$

The Wigner and Bohm potential are related through derivatives as

$$\rho\frac{\partial\ Q}{\partial\ x}=2\frac{\partial\ \rho\ Q_w}{\partial\ x}\ . \tag{49}$$

Below we write the quantum corrected momentum balance equation in terms of the Bohm potential. The details of this derivation are in Grubin *et al.* (1993). The equation is

$$\frac{\partial\ \rho(x)mv(x)}{\partial t}+\frac{\partial\ \rho(x)mv(x)^{\ 2}}{\partial x}= -\rho(x)\frac{\partial\left[V(x)+Q(x)/3\right]}{\partial x}-\frac{\partial\ \rho(x)k_BT}{\partial x}-\frac{\rho(x)mv(x)}{\tau}\ , \tag{50}$$

which is the standard form of the moment equation. In the absence of the Bohm potential, the classic equation is retrieved. As noted in an earlier section, the magnitude and sign of the quantum potential represent the quantum mechanical self-force on the particle; a consequence of which is the higher than classical value for the density within a barrier and a lower than classical value of density outside, but adjacent, to the barrier. This feature is represented under time independent zero current conditions, where the density, from (50) is given by

$$\rho(x) = N_C \exp-\left[\frac{V(x)+Q(x)/3-E_F}{k_B T}\right] . \tag{51}$$

The density as expressed by (51) provides a good qualitative and sometimes quantitative description of the quantum contributions to transport and is considered in detail in Grubin *et al.* (1993). The second moment equation is obtained by taking the second derivative of the quantum Liouville equation in the correlation direction. In the diagonal limit the second moment equation, using the definition of energy density, (15) yields for Fokker-Planck dissipation:

$$\frac{\partial E}{\partial t}+\frac{1}{2m^2}\frac{\partial P^{(3)}(x)}{\partial x}+\rho(x)v(x)\frac{\partial V(x)}{\partial x}+\frac{2E}{\tau}-\frac{4D}{m}\rho(x)=0 , \tag{52}$$

where $P^{(3)}(x)$, is the diagonal component of the matrix:

$$P^{(3)}(x,x') \equiv \left(\frac{\hbar}{2i}\right)^3 \frac{\partial^3 \rho(x,x')}{\partial x^3} . \tag{53}$$

Within the same approximate limit as that given for the quantum correction to the energy, we find

$$P^{(3)}(x) = \left[(mv)^2 + 3mk_B T + 4mQ_W\right]mv(x)\rho(x) , \tag{54}$$

and the second moment equation is

$$\frac{\partial E}{\partial t}+\frac{\partial\left[v(E+\rho k_B T\right]}{\partial x}= - \rho(x)v(x)\frac{\partial\left[V(x)+Q(x)/3\right]}{\partial x}-\frac{2}{3}\rho(x)Q_W\frac{\partial v}{\partial x}-\frac{2\left[E-E_0\right]}{\tau} . \tag{55}$$

In the above we have required that in the absence of an excitation the energy relax to some preassigned steady state value, E_0. This is guaranteed with $4D = 2mE_0/\tau$. A very similar discussion is given by Woolard *et al.* (1987).

In the above equations we note that the Bohm and Wigner quantum potentials appear with the factor '3'. Division by this multiplicative constant is the subject of some concern, as the single particle Schrodinger equation suggests a constant of unity [see (6)]. Some have taken the view (Ancona and Iafrate, 1989) that this is an adjustable parameter. A discussion in Grubin *et al.* (1994a) illustrates the effect of varying this parameter

There are several issues with respect to the quantum hydrodynamic equations. *First*, how well do they represent the space charge profiles? Side by side calculations have been performed and are discussed in Grubin *et al.* (1993). These illustrate that under certain circumstances the results using the quantum hydrodynamic equations and the quantum Liouville equation in the coordinate representation are virtually identical. *Second*, can the QHD provide quantitative evidence for negative differential resistance associated with resonant tunneling diodes. The work of Gardner (1994) suggests that this is possible. *Third*, can these equations be used in two dimensional simulations, and what are the consequences? Here the work of Zhou and Ferry (1992), and more recent work of Kreskovsky and Grubin (1994) indicate that the current voltage characteristics are only marginally altered, rather the distribution of charge is closer to what is expected from a full quantum treatment; i.e., there is more physics in the simulations. We now consider two examples of these studies.

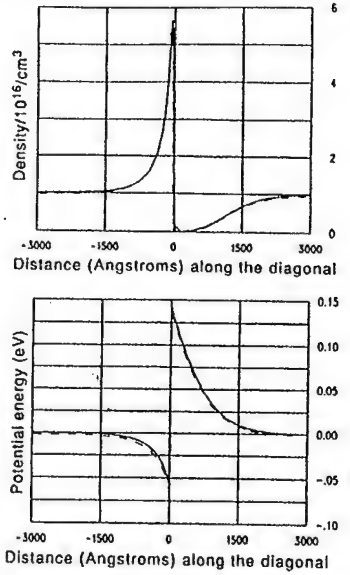


Figure 27. Density (a) and potential energy (b) distributions for a 200 meV diode, as obtained from the quantum Liouville equation in the coordinate representation (solid lines) and the quantum hydrodynamic equations (dashed lines) with the quantum potential divided by the factor '3'.

The first study compares the density and potential energy distributions for a 600 nm heterostructure with a constant background doping of $10^{22}/m^3$. A 200 meV abrupt barrier is placed across the right half of the structure. The self-consistent space charge and potential energy were computed for two values of applied bias, 0.0 eV and -0.2 eV. Boltzmann statistics were imposed for this study. Figure 27 displays the density and potential energy distribution corresponding to the flat band calculation. The results for the density matrix and

QHD equations are remarkably similar. The results for the bias of -0.2 eV are equally good.

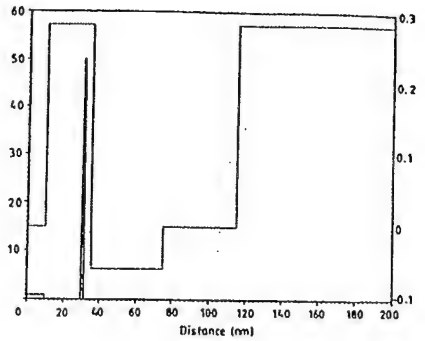


Figure 28. Heterostructure configuration of a quantum well FET for which the QHD equations were applied (dotted lines) Background density including delta doped region, for this calculation (Grubin et al., 1994a)

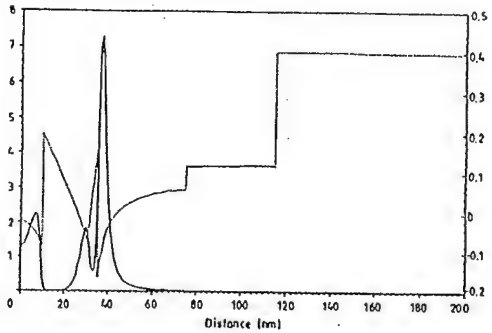


Figure 29. Self consistent distribution of potential energy (dotted lines) and density (solid lines) for the structure of Fig. 28.

A second example (Grubin *et al.*, 1994a) illustrates the use of the QHD equations in modeling more complex devices, such as a quantum well FET. Figure 28 displays the heterostructure configuration and the doping of a one dimensional slice through the source region of a quantum well FET (for the simulation the effective mass was constant). We se a barrier located 10 nm into the structure followed by a quantum well 20 nm long. A delta doped region placed within the wide barrier provides all of the carriers responsible for the operation of this structure. Figure 29 displays the self-consistent solution under flat band conditions. As seen from this latter figure there is a large concentration of carriers within the quantum well, a *with the peak in density occurring within the quantum well away from the barrier*. Note the approximately linear variation on either side of the delta doped region, with the consequent charge depletion.

It is important to note that in these studies the values of the quantum potential exceed the smallness criteria of a quantum correction. Rather the quantum contributions are of first order in magnitude!

ELECTRON AND HOLE TRANSPORT

Multispecies transport is part of the density matrix algorithm with applications to heterostructure barriers and diodes. The rather remarkable feature of the density matrix algorithm in the coordinate representation is the relative ease with which additional carries can be incorporated into the algorithm. Simple configurations such as *p-n* junctions and *n-p-n* structures have been examined as well as the more complicated heterostructure bipolar configuration, shown below, where quantum effects are expected to emerge at an abrupt wide band gap *n*-region. The calculations were performed within the context of Boltzmann statistics.

The structure we illustrate is an InAs HBT, 500nm long with an electron effective mass of 0.023 and hole effective mass of 0.4. The band gap was 400 meV. The HBT element consists respectively of (a) a 50 nm narrow band gap region (a donor concentration of $10^{22}/m^3$) followed by, (b) a 150 nm wide band gap emitter region with a conduction band offset of 150 meV, a valence band offset of 50 meV.(a donor concentration at $10^{22}/m^3$), (c) a 50 nm narrow band gap base (a acceptor concentration of $2.5\times10^{23}/m^3$) and (d) and a narrow band gap collector (with a $10^{22}/m^3$). The boundary conditions on the density were chosen so that the electron concentration was equal to background. The hole concentration on both the cathode and anode boundaries was chosen such that

$$n_{boundary}\, p_{boundary} = N_C\, N_V \exp-\left[E_G / k_B T\right] . \tag{56}$$

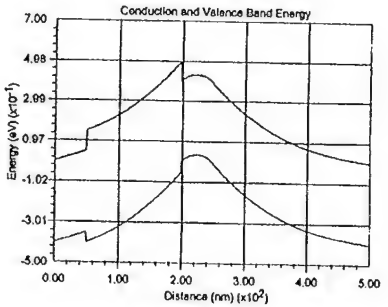


Figure 30. Conduction band profile for a InAs heterostructure bipolar transistor, obtained using solutions to the density matrix in the coordinate representation.

Figure 30 displays the conduction band profile for the HBT under equilibrium (Lake and Datta, 1992). The structure of the conduction band profile displays the familiar notch potential in going from the narrow band gap region to the wider band gap region. The

notch potential on the narrow band gap side signifies local depletion of charge from the wide band gap region to the narrow band gap region via real space transfer. The potential energy increases signifying that the carrier density is smaller than background. In the vicinity of the base, there is again a notch region indicating local accumulation of charge provided by the wide bandgap material. Any significant quantum effects are expected from this region (which if graded would diminish), although an examination of the quantum potential did not reveal any bound states. The distribution of electron density is displayed in Fig. 31, and the hole density is displayed in Fig. 32. For this calculation the electron density is everywhere less than that of background; whereas the hole density exceeds background everywhere except in the vicinity of the steep doping in the base.

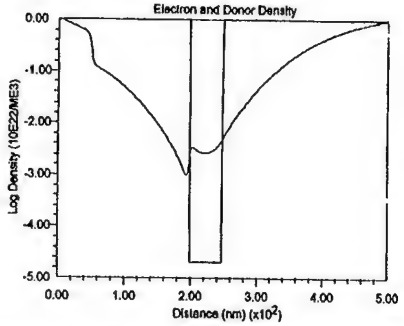


Figure 31. Electron density profile for the InAs heterostructure bipolar transistor of Fig. 30.

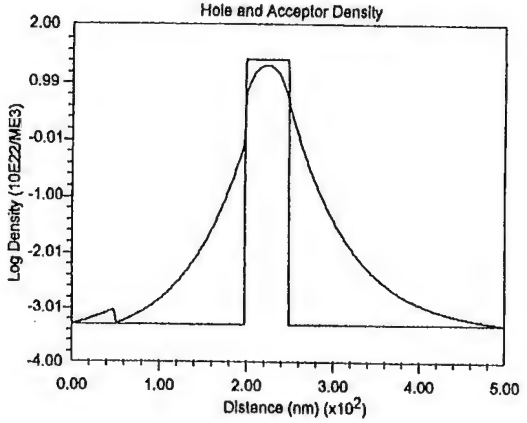


Figure 32. Hole density profile for the InAs heterostructure bipolar transistor of Fig. 30.

GREENS FUNCTIONS WITH APPLICATION TO ELECTRON DEVICES

The topics discussed in this series of lectures has focused on the application of the density matrix in the coordinate representation for the study of electron transport in semiconductor devices. There are, of course, alternative approaches, one of which is briefly discussed below.

Let us consider a simple generalization of equation (16) and introduce a 'two-time' density matrix defined as

$$\rho(x,x';t,t') = \sum_i f_i \Psi_i^*(x',t')\Psi_i(x,t) \quad . \tag{57}$$

A time-dependent equation of motion can be written down for the two time density matrix, although the form is not unique. For example with the following definitions:

$$\tau \equiv \frac{t+t'}{2}, \; \tau' \equiv \frac{t-t'}{2} \; , \tag{58}$$

we might deal with two equations. The first is

$$\left\{-i\hbar\frac{\partial}{\partial\tau} - \frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial x'^2}\right) + V(x,t) - V(x',t')\right\}\rho(x,x';t,t') = 0 \; , \tag{59}$$

where the Greek notation appears only in the time derivative. In the limit as $\tau' \to 0$, and zero scattering, (59) reduces to (11). The second equation is

$$\left\{-i\hbar\frac{\partial\,\rho}{\partial\,\tau'} - \frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial x'^2}\right) + V(x,t) + V(x',t')\right\}\rho(x,x';t,t') = 0 \; . \tag{60}$$

In discussing retarded Green's functions and its application to devices, apart from notational changes it is the right hand sides of (59) and (60) that are changed. For example from Ferry (1991) (equation 16.77) whose expression are generalized to three dimensions, the right hand sides of (59) and (60) become:

$$RHS(eq.59) \to \int dx''dt''\left[\Sigma(x,x'';t,t'')\,\rho(x'',x';t'',t') - \rho(x,x'';t,t'')\Sigma(x'',x';t'',t)\right] ,$$

$$RHS(eq.60) \to \int dx''dt''\left[\Sigma(x,x'';t,t'')\,\rho(x'',x';t'',t') + \rho(x,x'';t,t'')\Sigma(x'',x';t'',t)\right] + \frac{1}{2}\delta(\tau')\delta(x-x')$$

In the above, the term $\Sigma(x,x';t,t')$ represents nonlocal interactions such as that arising from phonons. Application of nonequilibrium Greens functions can be found in a number of studies, in particular Lake and Datta (1992).

Within the context of the above discussion, the density matrix in the coordinate representation as used in these lectures is an equal time correlation function. As discussed by

Datta *et al.*(1994), in steady state problems the correlation function depends only on the difference between the two times. With this in mind if we re-express the two time density matrix as a time difference density $\rho(x,x;\tau')$ and perform the following Fourier transform

$$\int d\tau' \rho(x,x';\tau')\exp-[i2E\tau'/\hbar] \equiv \rho(x,x';E) \ , \tag{61}$$

we can get detailed energy information from the density matrix. Thus, it appears easier to extract energy information from the two time density matrix (or Greens function) than from the equal time density matrix used in these lectures.

FUTURE DIRECTIONS

The density matrix in the coordinate representation provides a framework for studying transport in mesoscopic devices and in the more current technology- based devices such as HEMTs and HBTs. It should prove to be particularly valuable in addressing such key issues as the nature of the Schottky contact, the role of traps, and transient behavior. The algorithm used for solving the density matrix equation in the coordinate representation is based upon the method of characteristics which has the important feature of extremely rapid convergence times, and ease of use on workstations and PCs. The rapid convergence times are tied to equal-space meshes, which in turn tends to limit the size of devices that can be studied. Within this context, it appears that the approach discussed in these lectures is poised to become an advanced design tool for examining the physics and operation of quantum devices.

ACKNOWLEDGMENTS

All of the algorithm work was undertaken by T.R. Govindan (TRG) at Scientific Research Associates (SRA). Much of the interpretation of the results followed detailed discussion with TRG. The workstation interface for the simulation was developed by B. Morrison at SRA, and the PC interface was developed by Y. T. Chan at SRA. Parts of these lectures were excerpted from a forthcoming chapter. This work was supported by the Office of Naval Research, the Army Research Office and the Air Force Office of Scientific Research.

REFERENCES

Aharonov, Y., and Bohm, D., 1959, *Phys. Rev.* 115:485.
Ancona, M. A., and Iafrate, G. J., 1989, *Phys. Rev. B* 39:9536.
Bardeen, J., 1936, *Phys. Rev.* 49:653.
Bohm, D., and Hiley, B. J., 1993, "The Undivided Universe," Routledge, London.
Caldeira, A. O., and Leggett, A. J., 1983, *Physica* 121A:587.
Dekker, H., 1977, *Phys. Rev. A* 16:2126.
Eaves, L., Sheard, F. W., and Toombs, G. A., 1990, *in* "Physics of Quantum Electron Devices," Ed. by F. Capasso, Springer-Verlag, Berlin.
Ferry, D. K., 1989, *in* "Physics of Quantum Electron Devices," Ed. by C. Capasso, Springer-Verlag, Berlin, pp. 77.
Ferry, D. K., 1991, "Semiconductors," Macmillan, New York.
Ferry, D. K., and Grubin, H. L., 1993, *in* "Proc. Intern. Workshop on Computational Electronics," University of Leeds Press.
Ferry, D. K., and Grubin, H. L., 1994, in "Solid State Physics," Ed. by H. Ehrenreich, Academic Press, New York, in press.
Datta, S., Melloch, M. R., Bandyopadhyay, S., and Lundstrom, M. S., 1986, *Appl. Phys. Lett.* 48:487.
Datta, S., Klimeck, G., Lake, R., Lee, Y., and McLennan, M. J., 1994, *in* "Proc. Intern. Workshop on Computational Electronics," to be published.
De Groot, S. R., and Suttorp, L. G., 1972, "Foundations of Electrodynamics," North-Holland, Amsterdam.
Dirac, P. A. M., 1958, "The Principles of Quantum Mechanics," Oxford University Press, London.
Ehrenreich, H., and Cohen, M. H., 1959, *Phys. Rev.* 115:786.
Einevoll, G. T., and Sham, L. J., 1994, *Phys. Rev. B* 49:10533.
Frensley, W., 1990, *Rev. Mod. Phys.* 62:745.
Goldstone, J., and Gottfried, K., 1959, *Il Nuovo Cim.* 13:849.
Grubin, H. L., and Kreskovsky, J. P., 1989, *Sol. State Electron.* 32:1071.
Grubin, H. L., Govindan, T. R., Kreskovsky, J. P., and Stroscio, M. A., 1993, *Sol. State Electron.* 36:1697.
Grubin, H. L., Kreskovsky, J. P., Govindan, T. R., and Ferry, D. K., 1994a, *Semicond. Sci. Technol.* 9:855.
Grubin, H. L., Govindan, T. R., and Stroscio, M. A., 1994b, *Semicond. Sci. Technol.* 9:859.
Holland, P. R., 1993, "The Quantum Theory of Motion," Cambridge University Press, Cambridge.
Iafrate, G. J., Grubin, H. L., and Ferry, D. K., 1981, *J. de Physique* 10:C7-307.
Kluksdahl, N. C., Kriman, A. M., and Ferry, D. K., 1989, *Phys. Rev. B* 39:7720.
Krieger, J. B., and Iafrate, G. J., 1987, *Phys. Rev. B* 35:9644.
Krieger, J. B., and Iafrate, G. J., 1989, *Phys. Rev. B* 40:6144.
Kroemer, H., chien, W., Harris, J. S., Jr., and Edwall, D. D., 1980, *Appl. Phys. Lett.* 36:295.
Lake, R., and Datta, S., 1992a, *Phys. Rev. B* 45:6670.
Lake, R., and Datta, S., 1992b, *Phys. Rev. B* 46:4757.
Levine, F., Choi, K. K., Bethea, C. G., Walker, J., and Malik, R. J., 1987, *Appl. Phys. Lett.* 50:1092.
March, N. H., 1987, *in* "The Single-Particle Density of States in Physics and Chemistry," Ed. by N. H. March and B. M. Deb, Academic Press, London.
Messiah, A., 1961, "Quantum Mechanics," Vol. II, John Wiley, New York.
Philippidis, C., Bohm, D., and Kaye, R. D., 1982, *Il Nuovo Cim.* 71B:75.
Ravaioli, U., Osman, M. A., Pötz, W., Kluksdahl, N. C., and Ferry, D. K., 1985, *Physica B* 134:36.
Schroder, D. K., 1990, "Semiconductor Material and Device Characterization," John Wiley, New York.
Sollner, T. C. L. G., Goodhue, W. D., Tannewald, P. E., Parker, C. D., and Peck, D. D., 1983, *Appl. Phys. Lett.* 43:588.
Stroscio, M. A., 1986, *Microstruc. Superlatt.* 2:83.
Tsu, R., and Esaki, L., 1973, *Appl. Phys. Lett.* 22:562.

Weyl, H., 1927, *Z. Physik* 46:1.
Wigner, E. P., 1932, *Phys. Rev.* 40:749.
Woolard, D. L., Stroscio, M. A., Littlejohn, M. A., Trew, R. J., and Grubin, H. L.,1987, *in* "Proc. Intern. Workshop on Computations Electronics," Kluwer, Boston.
Zhou, J. R., and Ferry, D. K., 1992, *IEEE Trans. Electron Dev.* 39:1793.

SCREENING AND MANY-BODY EFFECTS IN LOW-DIMENSIONAL ELECTRON SYSTEMS

S. Das Sarma

Department of Physics
University of Maryland
College Park, Maryland 20742-4111 USA

INTRODUCTION

In this chapter, I will review the basic elements of many-body Green's function techniques as applied to one- and two-dimensional electron systems confined in ultrasmall semiconductor devices. In general, effects of many-body interaction are more pronounced in lower dimensions. With decreasing system size and increasing electron density, ultrasmall semiconductor devices should be viewed as interacting many-body mesoscopic quantum systems where electron-electron, electron-phonon, and electron-ionized impurity scattering processes all play fundamentally important roles, which often cannot be described by simple one electron theories. The main purpose of this chapter is to illustrate through some concrete examples how many-body Green's function techniques have been helpful in elucidating electronic properties of ultrasmall quantum devices. We will give particular attention to the role of dynamical screening in determining microscopic interaction effects in ultrasmall semiconductor devices.

This chapter has been organized in the following manner: We provide some basic formalism and theoretical preliminaries in the second section; in following, we deal respectively with dynamical screening and collective modes, and electron-electron interaction effects from a many-body theory perspective. In each chapter we discuss theoretical details as well as some concrete experimental situations where many-body effects are important. We conclude in section V by discussing several open questions with respect to the application of many-body theory to ultrasmall quantum devices and give several appropriate references where more detailed discussions on many-body theory (both for ultrasmall devices and for general formalism) can be found.

FORMALISM AND BACKGROUND

There are several excellent textbooks (Abrikosov *et al.*, 1963; Schrieffer, 1964; Fetter and Walecka, 1971; Mahan, 1981; Kadanoff and Baym, 1962) and review articles (Hedin and Lundqvist, 1969; Ando *et al.*, 1982) covering the basic aspects of (mostly) three-dimensional many-body theories. The subject has a long history (certainly longer than forty years, maybe even longer than sixty years depending on one's definition and taste). It is, of course, not possible to summarize or even mention all aspects of such an enormous subject in a set of lecture notes. This section on formalism is mainly intended to serve the purpose of introducing our definitions and notations -- we refer the reader to the cited books and reviews for the details on formalism.

All one electron properties of an interacting electron system can be calculated if the electron propagator or the single-particle Green's function, $G_{ij}(\mathbf{k},i\omega_n)$, is known. Here, ω_n is the finite temperature fermionic odd Matsubara frequency. The Green's function depends on the relevant (conserved) electron wavevector $\mathbf{k}$ which is one-dimensional (1D) or two-dimensional (2D) depending on the number of translationally invariant dimensions available in the problem. (We use effective mass approximation throughout our discussion of many-body theory, neglecting all lattice effects -- thus except for the confining potentials producing the ultrasmall device, the system is assumed to be homogeneous and translationally invariant in the dimensions other than the confinement direction with a conserved wavevector $\mathbf{k}$.) For quantum wires (or, other one-dimensionally confined structures) $\mathbf{k}$ is a 1D wavevector whereas for heterostructures, quantum wells, inversion layers, etc. $\mathbf{k}$ is a 2D wavevector. The subscripts (i,j) denote quantum subbands arising from the confining potentials. These quantum subbands have to be calculated (Stern and Das Sarma, 1984) from a suitable self-consistent solution of the corresponding noninteracting one electron confinement problem. In principle, the subband structure should be obtained from a coupled self-consistent solution of the one-electron Schrödinger equation and Poisson's equation in the Hartree (or, even better, in LDA) approximation. This calculation, by itself, could be a formidable numerical challenge for devices involving complicated confinement potentials. Most often in many-body calculations one makes a simple physical model for the noninteracting confinement problem (e.g., harmonic, rectangular, cylindrical, triangular confinement potentials) and uses an approximate analytical basis for the noninteracting subband wavefunctions in order to facilitate the calculation of many-body integrals.

The interacting many-body propagator $G_{ij}(\mathbf{k},i\omega_n)$ is related to the non-interacting free propagator $G^0_{ij}(\mathbf{k},i\omega_n)$ via the Dyson's equation

$$G^{-1}_{ij} = \left[G^0_{ij}\right]^{-1} - \Sigma_{ij} \ , \tag{1}$$

where $\Sigma_{ij}(\mathbf{k},i\omega_n)$ is the electron self-energy (also referred to as the mass operator) in the noninteracting subband basis. The self-energy operator defines the dressing of the noninteracting "effective mass" electrons by the relevant interaction perturbation. The renormalized or the dressed electron is referred to as a quasiparticle, which has an energy and a lifetime. Only in rather rare and specialized models can one calculate Σ exactly -- in most physical situations one has to resort to some renormalized perturbation expansion (or, some other suitable approximation) to calculate the self-energy. Since the noninteracting Green's function G^0_{ij} is exactly known, the interacting Green's function is easily calculable [from the Dyson's equation, i.e., our (1)] once Σ is obtained. The one-electron many-body problem, therefore, reduces to calculating the electron self-energy.

The noninteracting Green's function, $G^{v}_{ij}(\mathbf{k},i\omega_n)$, is, by definition, diagonal in the noninteracting (i,j) subband basis, with

$$\left.\begin{aligned} G^0_{ij}(\mathbf{k},i\omega_n) &= G^0_{ii}(\mathbf{k},i\omega_n)\delta_{ij} \\ G^0_{ii}(\mathbf{k},i\omega_n) &= \left[i\omega_n - E_i(k) + \mu\right]^{-1} \end{aligned}\right\} , \tag{2}$$

where δ_{ij} (= 1 if $i = j$, zero otherwise) is the Krönecker delta function, μ is the chemical potential of the system (at $T = 0$, $\mu = E_F$, the Fermi energy), and $E_i(k)$ is the noninteracting energy dispersion of the ith subband, which we assume to be parabolic (with no loss of generality, nonparabolicity can be easily included):

$$E_i(k) = E_i + \frac{k^2}{2m_i} . \tag{3}$$

In (3), E_i is the ith subband bottom energy (calculated in Hartree or some other mean field approximation scheme, or in a model approximation such as triangular, harmonic or square well confinement potential), and m_i is the effective noninteracting band mass of the ith subband. (We take $\hbar = k_B = 1$ throughout this chapter.) We assume $m \equiv m_i$ to be independent of i; if band nonparabolicity is quantitatively significant one replaces m by a subband mass $m_i(k)$, $m \equiv m_i(k)$, with no loss of generality. For finite temperature calculations within the grand canonical ensemble, the chemical potential μ, which is a parameter for the many-body calculation, should be determined at the end of the calculation by fixing the total electron density N:

$$N = \sum_i N_i , \tag{4}$$

where N_i is the electron density in the ith subband.

The electronic Hamiltonian for the ultrasmall semiconductor device is written as

$$H = H_0 + H_i , \tag{5}$$

where H_0 is the so-called noninteracting one electron part of the Hamiltonian, and all interaction effects are lumped into H_i. Note that, in addition to containing the electron kinetic energy and the confinement potential energy imposed by the appropriate external potentials (which produce the ultrasmall structure), the noninteracting Hamiltonian H_0 may contain certain average or mean-field part of the electron-electron interaction itself, as, for example, in the Hartree approximation where H_0 also contains the average Coulomb potential energy (obtained through a self-consistent solution of the Poisson's equation) arising from the electrons themselves. The one electron basis $\{i\}$ that diagonalizes H_0 is, by definition, the noninteracting subband basis

$$H_0|i\rangle = E_i|i\rangle . \tag{6}$$

The noninteracting Green's function G^0 is formally the standard resolvent operator:

$$G^0 = \left[z - H_0\right]^{-1} \tag{7}$$

in the noninteracting basis, where z is a complex frequency. The interacting Green's function G corresponding to the full Hamiltonian H is related to G^0 via the Dyson's equation, and is not diagonal in the noninteracting basis $\{i\}$ by virtue of the interaction H_i. Note that in order to avoid double-counting one must subtract out from H_i any average or mean-field part of the interaction (e.g., Hartree potential) which is already included in the noninteracting Hamiltonian H_0. Thus, G^0 contains all information about the noninteracting system.

Obtaining G now reduces to calculating the self-energy Σ in some approximation scheme, which we discuss in the subsequent chapters. Once G is known, all one electron quasiparticle properties can be explicitly calculated. For example, the quasiparticle density of states (D), the quasiparticle distribution function (n) the renormalized quasiparticle energy (E^*), the renormalized quasiparticle effective mass (m^*), the quasiparticle damping rate (γ), the electron lifetime (τ), the quasiparticle mean free path (ℓ), the subband bottom renormalization (Δ), and the quasiparticle renormalization factor (Z) are respectively given by

$$D_i(\omega) = \left(\frac{1}{2\pi}\right)\sum_{\mathbf{k}} A_i(\mathbf{k},\omega) \quad \text{per spin,} \tag{8}$$

$$n_i(k) = \int_{-\infty}^{0} \frac{d\omega}{2\pi} A_i(k,\omega) , \tag{9}$$

$$E_i^*(k) = E_i(k) + \Sigma_i\left(k, E_i^*(k) - \mu\right) , \tag{10}$$

$$m_i^*(k) = k\left[\frac{dE_i^*(k)}{dk}\right]^{-1} , \tag{11}$$

$$\gamma_i(k) = \left|\mathrm{Im}\,\Sigma_i\left(k, E_i^*(k) - \mu\right)\right| , \tag{12}$$

$$\tau_i(k) = \left[2\gamma_i(k)\right]^{-1} , \tag{13}$$

$$\ell_i(k) = k/\left[2m_i^*\gamma_i(k)\right] , \tag{14}$$

$$\Delta_i = E_i^*(k=0) - E_i = \mathrm{Re}\,\Sigma_i \ \ (\mathrm{k}=0) , \tag{15}$$

and

$$Z_i = \left[1 - \frac{\partial}{\partial\omega}\mathrm{Re}\,\Sigma_i(k,\omega)\right]^{-1}_{\omega = E_i^*(k) - \mu} . \tag{16}$$

In (8) to (16), we have introduced the subband quasiparticle spectral function $A_i(k,w)$:

$$A_i(k,\omega) \equiv -2\,\mathrm{Im}\,G_{ii}(k, i\omega_n \to \omega + i\eta), \ with\ \eta = 0+ . \tag{17}$$

We also note that the subband self-energy function is, in general, complex with

$$\Sigma_{ii}(k, i\omega_n \to \omega + i\eta) \equiv \mathrm{Re}\,\Sigma_i + i\,\mathrm{Im}\,\Sigma_i . \tag{18}$$

Here, $k \equiv |\mathbf{k}|$, and we assume isotropy so that all quantities depend only on the wavevector magnitude k. For brevity we often suppress the (k,ω) dependence.

The quasiparticle spectral function $A_i(k,\omega)$ is a central quantity in many-body considerations. Using Dyson's equation in (17) we see that

$$A_i(k,\omega) = \frac{-2\,\mathrm{Im}\,\Sigma(k,\omega)}{\left[\omega - E_i(k) + \mu - \mathrm{Re}\,\Sigma_i(k,\omega)\right]^2 + \left[\mathrm{Im}\,\Sigma_i(k,\omega)\right]^2} , \tag{19}$$

where $\omega \equiv \omega + i\eta$ with $\eta = 0+$ as a positive infinitesimal (the corresponding self-energy, $\Sigma(k,\omega)$, referred to as the retarded self-energy is obtained from the temperature self-energy $\Sigma(k,i\omega_n)$ by the analytic continuation $i\omega_n \rightarrow \omega + i\eta$). For noninteracting electrons, $\Sigma(k,\omega) \equiv 0$, and (19) reduces to

$$A_i(k,\omega) = 2\pi\,\delta[\omega - E_i(k) + \mu] , \tag{20}$$

with the δ-function in (20) signifying that the noninteracting energy $\omega+\mu$ must necessarily be $E_i(k)$. The presence of the self-energy function in (19) means that the interacting spectral function could be quite complicated which, in general [i.e., for arbitrary $\Sigma(k,\omega)$], may not have a δ-function piece at all. In the special situation where $\mathrm{Im}\,\Sigma = 0$ for a region of ω (for a fixed k) we obtain from (19), (10), and (16):

$$\begin{aligned} A_i(k,\omega) &= 2\pi\delta\left(\omega - E_i(k) + \mu - \mathrm{Re}\,\Sigma_i(k,\omega)\right) \\ &= 2\pi Z_i \delta\left(\omega - E_i^*(k) + \mu\right) \end{aligned} , \tag{21}$$

where Z_i, $E_i^*(k)$ are the quasiparticle renormalization factor and the quasiparticle energy, respectively. Thus, in the limit of vanishing $\mathrm{Im}\Sigma$, the interacting spectral function behaves similar to the noninteracting one with a quasiparticle peak at the "right" energy. For small $\mathrm{Im}\Sigma$, the interacting spectral function is approximately a Lorentzian (with the width determined by the damping rate γ) and the quasiparticle approximation is valid. The spectral function obeys the sum rule

$$\int_{-\infty}^{\infty} \frac{d\omega}{2\pi} A(k,\omega) = 1 , \tag{22}$$

which can be used to check a particular approximation scheme for the self-energy.

We emphasize that in general the self-energy and the interacting Green's function are off-diagonal in the subband representation because interaction causes (virtual) intersubband transitions. One should, therefore, diagonalize the self-energy (Σ_{ij}) and the Green's function (G_{ij}) matrices before calculating the physical quantities defined in (8)-(18). In practice, the diagonal approximation is extensively employed assuming that $G_{ij} = 0$ for $i \neq j$ even for the interacting Green's function. Explicit calculations show that the diagonal approximation is often justified in practice, because the off-diagonal components of the self-energy matrix are much smaller than the diagonal components $\Sigma_i \equiv \Sigma_{ii}$. But there are situations, particularly when resonance effects are important, where off-diagonal self-energy components cannot be neglected, and, in general, one should not neglect them without justification. We point out that the off-diagonal components of the self-energy directly depend on the off-diagonal components of the interaction matrix elements, which by virtue of the orthogonality of the subband basis {i}, vanish at long wavelengths and are comparable to the diagonal interaction matrix element only at large wavevectors ($k \sim a^{-1}$, where a is the approximate subband confinement size) when all components of the interaction matrix are generally small. This fact has been the rationale underlying the neglect of off-diagonal self-energy components in most theoretical calculations.

We have emphasized that a knowledge of the one electron interacting Green's function G allows us to calculate *all single-particle* properties of the interacting electron system. To obtain two-particle properties such as screening, conductivity, dynamical structure factor (which determines the scattering cross-section to external probes in the linear response theory), polarizability, etc. one must go beyond the one electron Green's function and calculate the corresponding interacting two-particle propagators (e.g., the current-current correlation function to obtain the 'conductivity, the density-density correlation function to obtain screening and polarizability). In general, a knowledge of G does not enable one to immediately calculate a two-particle propagator because of vertex corrections. In order to obey current conservation (and gauge invariance) one must satisfy certain relationships (Ward identities) between the self-energy and the vertex functions, and therefore vertex corrections cannot in general be neglected. (There are special situations where vertex corrections vanish and a knowledge of G then allows one to calculate the conductivity -- a well-known example being the zero-range isotropic delta function impurity scattering where $\gamma = |\mathrm{Im}\Sigma|$ directly gives the mobility scattering rate without any vertex correction.) We refrain from discussing vertex corrections in any details in these lectures for two reasons. One of the reasons is the need for brevity -- a discussion of vertex corrections will take far too long. The second, and in some sense a more important reason, is that most often in calculations of conductivity (or other two-particle properties) of ultrasmall devices one employs kinetic rate equations such as the Boltzmann equation where vertex corrections are implicitly incorporated. We, therefore, refer to the existing literature for details on vertex corrections.

Note that in our formal discussion so far we have not actually spelt out how to calculate the self-energy function (which then directly determines the interacting Green's function through the Dyson's equation) for a given interacting Hamiltonian. The actual calculation of the self-energy matrix depends crucially on the details of the interaction function H_i. For most calculations one employs a Feynman-Dyson perturbation theory in a renormalized interaction expansion. For Coulomb systems, which are of interest here, an expansion in the bare Coulomb interaction often diverges in each order of perturbation theory because of the highly singular (long-range) nature of Coulomb interaction. The trick, which was discovered (Quinn and Ferrell, 1958; Hedin, 1965; Rice, 1965) a long time ago, is to expand in a dynamically screened Coulomb interaction where the effect of long-range Coulomb interaction is included in the screening function. Formally this is a resummation of the perturbation series and one has to be careful to avoid double-counting. In the next section we discuss screening from a many-body perspective.

DYNAMICAL SCREENING

A central concept in the theory of interacting electron gas is dynamical screening (Pines and Nozieres, 1966), which renormalizes the bare Coulomb interaction between two electrons (or, for that matter, between an electron and an external point charge in the presence of many electrons) into a retarded (i.e., time dependent) non-local effective interaction. The dynamically screened interaction, $u(q, i\nu_m)$ where q is the wavevector and ν_m the bosonic even Matsubara frequency in the finite temperature formalism, is given by the following matrix equation:

$$u = v\varepsilon^{-1} , \tag{23}$$

where $v(q)$ is the bare Coulomb interaction and $\varepsilon(q, i\nu_m)$ is the dynamical dielectric function (also referred to as the screening function) for the electron system. Note that for a fully translationally invariant system (i.e., for a purely 1D, 2D, or 3D electron gas in the jellium effective mass approximation), Eq. (23) represents an ordinary scalar algebraic equation which completely determines the screened interaction u once the dielectric function is known -- $v(q)$ being known by definition since it is the Fourier transform of the $1/r$ Coulomb interaction in the appropriate dimension. In ultrasmall quantum devices the system is not homogeneous in the direction(s) of confinement and $\mathbf{q}$, as usual, denotes a wavevector in the translationally invariant dimensions perpendicular to the confinement direction. Because of this loss of translational invariance arising from confinement potential, (23) represents a (fourth-rank) tensor screening equation for ultrasmall devices with u, v, ε being matrices which need to be evaluated in the appropriate subband basis set $\{I\}$:

$$u_{ijlm} \equiv \langle lm|u|ij\rangle \ , \tag{24}$$

$$v_{ijlm} \equiv \langle lm|v|ij\rangle \ , \tag{25}$$

$$\varepsilon_{ijlm} \equiv \langle lm|\varepsilon|ij\rangle \ , \tag{26}$$

where $|i\rangle$, etc. are the self-consistent (noninteracting) subband wavefunctions discussed in the previous section. All the components of the dynamically screened interaction u can be calculated from (23)–(26) once the dielectric matrix ε is explicitly known. In the extreme quantum limit, when only the lowest quantum subband of the ultrasmall device is occupied by the carriers, the one subband approximation is often employed for the sake of simplicity, keeping only the ground subband in (24)–(26). Then, (23) becomes a scalar equation with u, v, ε being evaluated as matrix elements in the ground subband.

Over the last forty years or so a great deal of work has gone into an accurate evaluation of the dynamical dielectric function of an interacting electron gas. In general, this is a formidable problem even in a translationally invariant bulk 3D electron gas and inclusion of electron-electron interaction effect into the dielectric function can only be done approximately. For the ultrasmall devices of our interest, the task is even more daunting because the loss of translational invariance formally converts the problem into a many-body interacting multicomponent electron system (with each subband i representing a different component) whose dielectric matrix is extremely complicated. Fortunately, a simple self-consistent mean-field approximation, which includes only the long-range Coulomb interaction in the dielectric response leaving out all exchange-correlation corrections, seems to work extremely well for most practical purposes. This is the celebrated random-phase-approximation (RPA), which is just the time-dependent Hartree approximation (i.e., includes only Coulomb interaction through Poisson's equation, neglecting all quantum fluctuations). RPA is known to be exact at long wavelengths ($\mathbf{q} \to 0$) and high electron densities ($r_s \to 0$, where r_s, the average inter-particle distance measured in the units of effective Bohr radius is the standard electron gas parameter signifying the strength of many-body electron-electron interaction effect) by virtue of sum rules arising from particle conservation (the f-sum rule) and the long wavelength ($\mathbf{q} \to 0$) divergence of the bare Coulomb interaction. Systematic improvements on RPA which incorporate vertex corrections in various approximations are possible, but will not be discussed much in this article because of their highly technical nature.

The dielectric function ε is simply related to the density-density correlation function (and is, therefore, a two-particle propagator), which in the many-body language is the improper or the reducible polarizability function. In particular, the reducible polarizability function, $\tilde{\pi}(q, i\nu_m)$, is given by

$$\tilde{\pi} = \pi\varepsilon^{-1} \ , \tag{27}$$

where π is the irreducible polarizability function. The reducible polarizability (or, equivalently, the density correlation) function obeys a Dyson's equation for two-particle propagators given by

$$\tilde{\pi} = \pi + \pi v \tilde{\pi} \ , \tag{28}$$

where v is the bare Coulomb interaction. This immediately gives

$$\tilde{\pi} = \pi(\mathbf{1} - v\pi)^{-1} \ , \tag{29}$$

making [comparing with (27)],

$$\varepsilon = \mathbf{1} - v\pi \ . \tag{30}$$

Thus, ε is known once the irreducible two-particle propagator π is known. As discussed before, a full knowledge of the one electron Green's function G is, by itself, not sufficient to obtain the irreducible polarizability π because of the vertex function Γ which must also be known. Formally,

$$\pi = GG\Gamma \ , \tag{31}$$

with the vertex function Γ obeying the formal functional derivative relationship (Ward identity):

$$\Gamma = -\frac{\delta G^{-1}}{\delta v} \equiv \mathbf{1} + \frac{\delta \Sigma}{\delta v} \ . \tag{32}$$

Vertex correction is formally neglected in RPA, i.e., $\Gamma = \mathbf{1}$ which means that the noninteracting Green's function G^0 (which contains the confinement and self-consistent field potentials) can then be used to calculate the RPA polarizability $\pi^0 \equiv G^0G^0$, which becomes (note that $G^0_{ij} = G^0_{ii}d_{ij}$):

$$\pi^0_{ij}(\mathbf{q}, i\nu_m) = -g\sum_{\mathbf{k}} \frac{n_i^0(\mathbf{k}+\mathbf{q}) - n_j^0(\mathbf{k})}{i\nu_m - E_i(\mathbf{k}+\mathbf{q}) + E_j(\mathbf{k})} \ , \tag{33}$$

where $n_i(\mathbf{k})$ is the noninteracting Fermi distribution function in the i*th* subband given by

$$n_i(\mathbf{k}) = \left[1 + \exp\{E_i(k) - \mu\}/T\right]^{-1} \ , \tag{34}$$

and g is a degeneracy factor (=2, for spin half electrons in a single valley) which includes both spin and valley degeneracies. Because (33) contains only noninteracting energy whose wavevector dispersion is exactly known, the integration over $\mathbf{k}$ can be analytically

performed at $T = 0$ to give the noninteracting polarizability function [the so-called Lindhard (1954) screening function] in the relevant dimensions. The explicit forms for the RPA polarizability in 1D, 2D, and 3D electron systems can be found in the literature (Lindhard, 1954; Stern, 1967; Jain and Das Sarma, 1987, 1988; Li and Das Sarma, 1991). Once the screening function π is known the dielectric function is given by $\varepsilon = 1 - v\pi$.

While the static dielectric function $\varepsilon(q,iv_m=0)$ gives the statically screened Coulomb interaction $u(q,0) = v(q)\varepsilon(q,0)^{-1}$, the zeros of the dielectric function define the collective plasma modes (i.e., charge density excitations) of the system. Since the subband dielectric function for the ultrasmall device is a tensor given by

$$\varepsilon_{ij\ell m}(q,\omega) = \delta_{i\ell}\delta_{jm} - v_{ij\ell m}\pi_{\ell m} \ , \tag{35}$$

the collective charge density excitations of the device are given by the determinental equation

$$\left|\varepsilon_{ij\ell m}(q,\omega)\right| = 0 \ . \tag{36}$$

In (35) and (36), $iv_m = \omega + i\eta$ is the usual analytic continuation. Note that the solutions, $\omega \equiv \omega(q)$, of (36), in general, define all the collective charge density excitations of the system. The modes with $i = j = \ell = m$ are the intrasubband plasma oscillations whereas the modes with different i, j, ℓ, m are intersubband collective modes. These modes have been studied extensively (Das Sarma, 1991) in ultrasmall structures both theoretically and experimentally via inelastic light scattering and far infrared absorption spectroscopies.

ELECTRON-ELECTRON INTERACTION

Dynamical screening and collective mode behavior discussed in the last section are only two specific features of electron-electron interaction associated with the linear response of the system. In the RPA, which is extensively employed, only the long-range Coulomb interaction between electrons is included in the theory (in a self-consistent time-dependent Hartree approximation) leaving out all quantum fluctuations associated with short-range exchange-correlation effects. Inclusion of exchange-correlation effects in the dynamical screening properties requires incorporation of vertex corrections in the theory which, as emphasized before, must be done obeying the Ward identities in order to maintain current conservation.

One popular method of incorporating vertex correction in screening is to try to sum all the ladder diagrams in the irreducible polarizability function, thus incorporating excitonic electron-hole interaction in the theory. If one assumes that this excitonic electron-hole interaction is short-ranged (i.e., δ-function like in real space), one can exactly sum the geometric series representing the ladder diagrams to obtain

$$\pi = \pi^0(1 + \tilde{u}\pi^0)^{-1} \ , \tag{37}$$

where π is the vertex-corrected irreducible polarizability including all ladder diagrams, π^0 is the leading-order polarizability bubble without any vertex correction, and $\tilde{u}$ is the short-ranged effective electron-electron interaction in the wavevector space (which is sometimes referred to as the exchange interaction in this context.) The reducible polarizability or, equivalently, the dielectric function is then given by the usual sum of all the bubbles formed by π to obtain

$$\varepsilon = 1 - v\pi = 1 - v\pi^0(1 + \tilde{u}\pi^0)^{-1} \ , \tag{38}$$

where v is the usual (unscreened) direct Coulomb interaction.

An important question is what functional form to use for the vertex interaction function $\tilde{u}$. Note that while $\tilde{u}$ could be a renormalized screened interaction, v in (38) is the unscreened bare Coulomb interaction. Note also that if one uses the bare Coulomb interaction v itself as the vertex interaction in the ladder series, as would be appropriate for a bare exciton in a very dilute electron-hole system where screening by other electrons is unimportant, then the full vertex corrected polarizability function can only be found by solving an integral equation (the so-called Bethe-Salpeter equation) representing the ladder diagrams. The ladder diagram series can be explicitly summed for an effective short-range interaction which converts the integral equation to an algebraic geometric series by virtue of the delta function kernel in the integral series. We should point out that the approximation of keeping all the ladder diagrams in the irreducible polarizability is sometimes referred to as the generalized random phase approximation (GRPA) or random phase approximation with exchange (RPAE). If one includes the exchange self-energy in G, i.e., calculates G in the Hartree-Fock approximation then this ladder-bubble approximation for the polarizability is the time dependent Hartree-Fock approximation (TDHF) in the sense RPA is just the time dependent Hartree approximation.

A very useful alternative to the diagrammatic many-body theory for dynamical screening in inhomogeneous systems is to use the density functional theory where one deals with effective one-electron Schrödinger-like Kohn-Sham equations in a non-local potential. A simple local approximation to this non-linear self-consistent scheme, the so-called LDA (local density approximation), seems to work extremely well (both in general and for ultrasmall devices in particular). The time-dependent version of the LDA theory, TDLDA, has been extensively used (Marmorkas and Das Sarma, 1993) to study dynamical response of ultrasmall quantum structures. Results of TDLDA calculation are in impressive agreement with experiments, and we provide a discussion of TDLDA in the appendix of this article.

Once the dynamically screened Coulomb interaction, $u(q,iv_m)$, has been calculated the electron self-energy arising from electron-electron interaction effects is usually obtained in a perturbative series expansion in u, and the leading term (formally Gu) in this series is one of the most extensively used non-trivial approximation (Quinn and Ferrell, 1958; Hedin, 1965; Rice, 1965) for the electron self-energy. In this leading dynamically screened Coulomb interaction approximation the subband self-energy matrix for the ultrasmall device has the following form:

$$\Sigma_{ij}(k,i\omega_n) = (-T)\sum_{ij}\int\frac{dq}{(2\pi)^d}\sum_{iv_m}u_{ijij}(q,iv_m)G_{ij}(k-q,i\omega_n - iv_m) \ , \tag{39}$$

where d is the dimensionality of the device and T is the temperature. This leading-order (in dynamically screened interaction u) electron-electron self-energy expression, which has been used extensively in both bulk and low-dimensional systems, is often referred to as the "GW approximation" in the literature. The nomenclature derives from the fact that functionally the self-energy in (39) is given by $\Sigma \sim Gu$, and the notation W is often used in the literature for the dynamically screened interaction u, making $\Sigma \sim GW$, and hence the terminology GW approximation. In spite of its apparently leading-order perturbative nature, the GW approximation [i.e., (39)] is really the best systematic theory for electron self-energy available in the literature, and overall its agreement with experiments in 3D (i.e., metals), 2D

(i.e., inversion layers, heterostructures, and quantum wells), and 1D (i.e., quantum wires) is very impressive. It should be emphasized that in spite of its formal leading-order perturbative form, the self-energy function defined by (39), in fact, contains infinite order terms in bare interaction because the dynamically screened interaction u is an infinite geometric series in v

$$u = v\varepsilon^{-1} = v(\mathbf{1} - v\pi)^{-1}$$

$$= v + v\,\pi\,v + v\,\pi\,v\,\pi v + \ldots \,. \tag{40}$$

Substituting (40) in (39), one easily sees that in terms of the bare Coulomb interaction v the "GW" approximation to the self-energy represents an infinite order perturbation series. The first term in this series, formally given by Gv where v is the bare Coulomb interaction, is the exchange self-energy Σ or the Hartree-Fock self-energy $\Sigma_{HF} \equiv \Sigma_x$. Note that the exchange self-energy is frequency-independent by virtue of the frequency independence of v. The rest of the self-energy (i.e., $\Sigma - \Sigma_x$) is sometimes called the correlation self-energy. It is worthwhile to point out that keeping just Σ_x ($\sim Gv$) in the calculation neglecting all the correlation contributions is generally a bad approximation. Clearly all vertex corrections are neglected in the "*GW* approximation" of the self-energy (i.e., $\Gamma = \mathbf{1}$), and systematic improvements are, in principle, possible (Marmorkos and Das Sarma, 1991) by incorporating vertex corrections in the self-energy in the "*GWΓ*" approximation.

The calculation of the finite-temperature self-energy of confined electrons in ultrasmall semiconductor structures, even within the leading-order "*GW* approximation" as defined by (39), is a thoroughly non-trivial task and only a few attempts (Das Sarma and Vinter, 1982) have been made at full evaluations of (39) in the subband basis. The following additional approximations (i.e., in addition to the neglect of vertex corrections) are often made in evaluating Σ as defined by (39):

(1) The Green's function G_{ij} appearing inside the integrals on the right hand side is taken to be the zeroth order noninteracting propagator G^0 which is diagonal in the subband indices and, therefore, (39) becomes

$$\Sigma_{ij}(k,i\omega_n) = (-T)\sum_{i'}\sum_{\mathbf{q}}\sum_{iv_m} u_{iii'j'}(q,iv_m)G^0_{i'i'}(\mathbf{k}-\mathbf{q},i\omega_n - iv_m)\,, \tag{41}$$

where, as usual,

$$\sum_{\mathbf{q}} \equiv \int \frac{d\mathbf{q}}{(2\pi)^d}\,.$$

This gets rid of the rather formidable problem of self-consistency between G and $\Sigma = \int Gu$. Note that repeated iteration with successively higher-order G's can improve upon this approximation. We know of no such self-consistent "*GW*" calculation for ultrasmall devices in the electron-electron interaction context even though the corresponding approximation for the electron-impurity self-energy, referred to as the self-consistent Born approximation, has been carried out (Das Sarma and Vinter, 1981; Das Sarma and Xie, 1987, 1988; Hu and Das Sarma, 1994).

(2) The self-energy is calculated (Vinter, 1976, 1977) in the $T = 0$ limit when it simplifies somewhat.

(3) Subband diagonal approximation is made by neglecting the off-diagonal elements of the interaction matrix elements and assuming that

$$u_{ij\ell m} \propto \delta_{ij}\delta_{\ell m} \tag{42}$$

which makes the self-energy appearing in (49) purely diagonal in the subband approximation:

$$\Sigma_{ii}(k,i\omega_n) = (-T)\sum_{iv_m} u_{iiii}(q,iv_m)G_{ii}(\mathbf{k}-\mathbf{q},i\omega_n - iv_m)\,. \tag{43}$$

We emphasize that this approximation is unnecessary and needs to be justified through detailed calculations that the dynamically screened interaction indeed satisfies the diagonal approximation of (42).

(4) Additional approximations (beyond RPA) are made in evaluating the integrals in Eq. (39). One popular approximation (Vinter, 1976, 1977) is to neglect the full frequency dependence of the RPA dielectric function and to approximate ε^{-1} as a collection of plasmon poles at the effective collective mode frequencies. This approximation, which enables one to do the frequency sum in (39) rather trivially, is called the plasmon-pole approximation.

We now discuss a recent calculation (Hu and Das Sarma, 1992, 1994) of the lowest subband self-energy based on (43) for electrons confined in the conduction band of a GaAs quantum wire structure. The calculation makes the one subband approximation keeping only the ground subband $i = 1$ and neglecting all other subbands. Then, (43) can be rewritten as

$$\Sigma_{11}(k,i\omega_n) = -T\int \frac{d^d\mathbf{q}}{(2\pi)^d}\sum_{iv_m} u_{1111}(q,iv_m)G^0_{11}(\mathbf{k}-\mathbf{q},i\omega_n - iv_m)\,. \tag{44}$$

Using the spectral representation for the interaction operator u and doing the usual analytic continuation $i\omega_n \to \omega + i\eta$ the retarded self-energy can be written as

$$\Sigma(k,\omega) = \int \frac{d\mathbf{q}}{(2\pi)^d}\int_{-\infty}^{\infty}\frac{d\omega'}{2\pi}\frac{B(q,\omega')}{\omega+\omega'-E_1(\mathbf{k}-\mathbf{q})-i\eta}\times\left[n_B(\omega')+n_F(E_1(\mathbf{k}-\mathbf{q})-\mu)\right]\,, \tag{45}$$

where

$$B(q,\omega') = 2\,\mathrm{Re}\,u(q,iv_m \to \omega + i\eta) = -i[u(q,\omega+i\eta) - u(q,\omega-i\eta)]\,, \tag{46}$$

and

$$n_F(x) = \left[1+e^{x/T}\right]^{-1}$$
$$n_B(x) = \left[-1+e^{x/T}\right]^{-1}\,. \tag{47}$$

In (45)-(47) we use the abbreviation $\Sigma \equiv \Sigma_{11}$; $u \equiv u_{1111}$. The remaining integrals are now done remembering that

$$u = v\varepsilon^{-1} \tag{48}$$

$\varepsilon = 1 - v\pi^0$ in RPA with the RPA 1D polarizability at finite temperatures being given by

$$\pi^0(q,\omega,T,\mu) = \int_0^{e^{\mu/T}} \frac{dx}{(x+1)^2} \pi^0(q,\omega,T=0,\mu - T\ln(x)) \text{ for } \mu < 0 \,, \qquad (49)$$

and,

$$\pi^0(q,\omega,T,\mu) = \int_{e^{-\mu/T}}^{\infty} \frac{dx}{(x+1)^2} \pi^0(q,\omega,T=0,\mu + T\ln(x)) + \int_0^{\infty} \frac{dx}{(x+1)^2} \pi^0(q,\omega,T=0,\mu - T\ln(x)) \text{ for } \mu > 0. \qquad (50)$$

The zero-temperature noninteracting 1D polarizability (the 1D Lindhard function) $\pi^0(q,\omega,T=0,\mu)$ is given by the analytic formula

$$\pi^0(q,\omega) = \frac{m}{\pi q} \ln\left\{ \frac{\omega^2 - \left[(q^2/2m) - qv_F\right]^2}{\omega^2 - \left[(q^2/2m) + qv_F\right]^2} \right\} \,, \qquad (51)$$

where the principal value of the logarithm ($|\mathrm{Im}[\ln] < \pi|$) should be taken. In (51), v_F is the 1D Fermi velocity given by

$$\mu(T=0) = E_F = \tfrac{1}{2} m v_F^2 \,. \qquad (52)$$

A very detailed numerical calculation of the finite temperature self-energy for 1D quantum wires has recently been carried out (Hu and Das Sarma, 1992, 1993).

Before concluding this section we briefly discuss the behavior of the quasiparticle spectral function $A(k,\omega)$ near the Fermi surface. To do this we need to know the real and the imaginary parts of electron self-energy for small $|\omega|$. Calculation of $\Sigma(k_F,\omega)$ can be analytically done in $d = 1$, 2, and 3, and the results are (for $|\omega| \to 0$):

$$\mathrm{Re}\,\Sigma \sim \begin{cases} \omega \,, & d=3, \\ \omega \,, & d=2, \\ \omega \ln|\omega| \,, & d=1, \end{cases} \,, \qquad (53)$$

and

$$\mathrm{Im}\,\Sigma \sim \begin{cases} \omega^2 \,, & d=3, \\ \omega^2 \,, & d=2, \\ |\omega|\sqrt{\ln|\omega|} \,, & d=1. \end{cases} \qquad (54)$$

It follows immediately from (53) and (54) that the electron spectral function near the Fermi surface, $A(k_F,\omega\to0)$, has a δ-function peak at $\omega = 0$ for both 3D and 2D systems, *but not for 1D systems*. In particular, recalling the definition of the spectral function from (19), and using (53) and (54) we find that

$$A(k_F,\omega\to0) \sim \delta(\omega - E(k_F) + \mu) = Z_F \delta(\omega) \quad \text{for } d = 2,3, \qquad \sim \left(|\omega| \left|\ln(|\omega|)\right|^{\gamma_2}\right)^{-1} \quad \text{for } d = 1. \qquad (55)$$

This leads to the rather startling conclusion that while in $d = 2$ and 3, the interacting electron system allows for the existence of well-defined quasiparticles close to the Fermi surface, no such quasiparticle can exist in $d = 1$ because the interacting spectral function has no δ-function piece by virtue of the slow fall-off of $\mathrm{Im}\Sigma(\omega\to0)$. An equivalent way of stating this result is that the quasiparticle renormalization factor Z_F at the Fermi surface is finite in $d = 2$ and 3, and it vanishes for $d = 1$ leading to a continuous momentum distribution function n(k) in a one-dimensional interacting electron gas with the consequent vanishing of the Fermi surface. A profound conceptual consequence of this behavior is that strictly there is no Landau Fermi liquid theory in one-dimension -- the system should be treated as a manifestly strongly interacting system with no one to one correspondence with the noninteracting system. Such a system is referred (Haldane, 1981; Sólyom, 1979) to as a Luttinger liquid in contrast to 2D and 3D electron systems which are Fermi liquids. While we have discussed these results here based purely on the leading-order "*GW* approximation," the qualitative conclusion that quasiparticles exist in two-and three-dimensional electron systems but *not* in one dimension (or, equivalently, the interacting momentum distribution function has a discontinuity at $k = k_F$ for $d = 2$ and 3 but is continuous for $d = 1$) is believed to be valid in general. (The detailed form of the 1D self-energy depends on the "*GW* approximation," however.) The extent to which device performance and device simulation in ultrasmall quantum wire devices may depend on this fundamental result is, however, unclear (Hu and Das Sarma, 1992, 1993). We believe that for most (if not all) practical purposes of device physics the Luttinger liquid aspects of a quantum wire structure can be ignored for the following reasons:

(1) Both finite temperature and finite impurity scattering effects severely suppress this behavior.
(2) The effect, while being conceptually profound, is quantitatively quite small.
(3) There is obviously no Fermi liquid type behavior in an interacting system even for $d = 2$ and 3 far away from the Fermi surface (i.e., $|\omega| \neq 0$) and for high-field device performance the region around the Fermi surface is not necessarily the most significant region.

Before concluding this section we mention that, with suitable re-definitions of the interaction function u entering into the self-energy calculations [(39)-(43)], the formalism discussed here can easily be used for the evaluation of electron-phonon and electron-impurity self-energies. For the calculation (Jalabert and Das Sarma, 1989; Das Sarma *et al.*, 1990a) of electron-phonon self-energy, u should be the total effective dynamically screened interaction between two electrons including effects of both the direct Coulomb interaction and the interaction mediated by virtual phonon exchange. For electron-impurity self-energy calculations (Das Sarma and Vinter, 1981; Das Sarm and Xie, 1987, 1988; Hu and Das Sarma, 1994), u is the screened interaction between an electron and a static impurity. One has the additional requirement of an ensemble averaging for the electron-impurity interaction problem because of the presence of disorder. Specific examples of such electron-phonon and electron-impurity self-energy calculations in ultrasmall semiconductor devices can be found in the literature (Das Sarma and Vinter, 1981; Das Sarm and Xie, 1987, 1988; Hu and Das Sarma, 1994; Jalabert and Das Sarma, 1989; Das Sarma *et al.*, 1990a).

CONCLUSION

Ultrasmall semiconductor devices are particularly difficult to deal with as interacting many-body systems because the loss of translational invariance associated with confinement considerably complicates the relevant many-body calculations. As emphasized throughout

this article, carrier confinement presents two particular difficulties in many-body calculations compared with the text-book examples of bulk 3D jellium theories. The first problem is that all calculations must be carried out in suitable basis of confined one-particle wavefunctions. In general, the one-electron confined wavefunctions must be calculated using a fully self-consistent scheme where (at least) the Hartree potential is included. This means that all the interaction matrix elements entering (e.g., u_{ijlm}) the many-body theory can only be obtained numerically only after tedious self-consistent calculations. The subsequent many-body calculation is, therefore, extremely demanding, and we know of only one finite temperature electron self-energy calculation (Das Sarma and Vinter, 1982) in low-dimensional semiconductor structures where such a full theory using the Hartree basis functions was carried out. Most theories parametrize the confinement wavefunctions using simple analytical confinement potentials, and typically use drastic approximations in the starting many-body Hamiltonian. An associated difficulty is the matrix nature of the many-body self-energy and the interacting Green's function which makes the problem equivalent to doing many-body calculations for a strongly interacting multicomponent electron system. One often resorts to diagonal approximation to avoid this complication. We should emphasize that in calculating (Jalabert and Das Sarma, 1989) the self-energy for electron-phonon interaction in ultrasmall structures the inhomogeneity problem associated with confinement is even more difficult because in addition to electron confinement effects one must now include effects of confinement on phonons as well in the starting many-body Hamiltonian. This task has been attempted in very few publications with most many-body electron-phonon calculations making the bulk phonon approximation neglecting all phonon confinement effects.

The second problem associated with many-body effects in ultrasmall structures is more subtle and is related to the low-dimensional nature of these devices. Many-body approximations become progressively worse as the system dimensionality is lowered because the relative significance of interaction is enhanced in lower dimensions due to phase-space restrictions. An extreme example of this is, of course, the Luttingher liquid behavior of one-dimensional electron systems where any finite electron-electron interaction is manifestly a nonperturbative effect, driving the system into a non-Fermi liquid like behavior. In 2D systems, quantitative many-body effects are relatively more important than in corresponding 3D systems. There are two consequences: Inclusion of many-body effects in calculations becomes more necessary in lower dimensional systems, and low-order perturbation expansions (e.g., the "*GW* approximation" for the self-energy) become less reliable. Since it is, in general, very difficult, if not impossible, to do anything better than "*GW*"-type leading order screened interaction calculations in ultrasmall structures, one should bear in mind the possible lack of relative reliability of such many-body calculations in low dimensional systems.

Before concluding it may be worthwhile to briefly discuss the regime of doping or carrier densities where many-body electron-electron interaction effects could be significant in ultrasmall semiconductor devices. The standard (zero-temperature) many-body parameter for an interacting electron system is the r_s-parameter with r_s being the dimensionless average inter-electron distance measured in units of effective Bohr radius. High electron density implies low r_s and vice versa. The noninteracting energy of an electron system being all kinetic energy goes as r_s^{-2} whereas the leading behavior of the interaction energy must be r_s^{-1} on dimensional grounds since the Coulomb interaction is $1/r$. Thus many-body effects are important for large r_s with the non-interacting kinetic energy dominating the small r_s behavior. If we arbitrarily choose the $r_s = 1$ value as the dividing point between strongly and weakly interacting systems (we emphasize that this choice is entirely arbitrary), then one concludes that for ultrasmall 2D confined structures many-body effects become particularly important for the 2D electron density less than 10^{11} (10^{12}) cm^{-2} for GaAs(Si) based devices. (The effective r_s-parameter is substantially larger in Si than in GaAs due to the much smaller value of the effective Bohr radius in Si.) For comparison the 3D metals have $r_s \sim 2$–6 range. For 1D GaAs structures, if we ignore the Luttinger liquid complications, then $r_s \sim 1$ around an electron density of 10^6 cm^{-1}. We emphasize, however, that finite temperature introduces a new dimensionaless parameter T/T_F where T_F is the Fermi temperature, and for $T/T_F > 1$, many-body effects are usually small.

We now conclude with the mot important question (for our purpose): How much should one worry about many-body effects (and dynamical screening) in considering ultrasmall semiconductor quantum device operation? No unique answer to this question really exists, and the answer obviously depends a great deal on the context and on the parameter range of operation of the particular device. There is one aspect, however, where an answer to a part of this question can be given quite unequivocally. Dynamical screening should *always* be included (at least within the RPA) in considering any interaction or scattering process in ultrasmall devices. There may be situations (particularly at high electron densities) where static screening (i.e., putting $i\nu_m = 0$ in ε) may be adequate, but in general static screening overestimates (Das Sarma *et al.*, 1988a, 1988b, 1990b) the strength of screening. The usual dynamical screening effect typically lies in between the static screening and (unscreened) bare interaction results although at low electron densities it is possible to see (Das Sarma *et al.*, 1988a, 1988b, 1990b) a small anti-screening effect. (The anti-screening effect of dynamical screening is never very large, though.) With respect to many-body renormalization effects it should be remembered that ultrasmall semiconductor devices are usually made of ultrapure materials which tend to enhance interaction effects. With increasing computer power it is probably reasonable to start including at least some aspects of quasiparticle renormalization effects in quantum device simulation codes. While it may be difficult to do so computationally at the present time, perhaps in the next five years increasing computing power of available workstations will make this possible. Even if one does not include full effects of many-body quantum fluctuations in device modelling, one should perhaps incorporate exchange-correlation effects within the mean field approach by using the computationally more tractable LDA and dynamical TDLDA theories (see Appendix). Finally, we mention that all our discussions and many-body formalism for ultasmall semiconductor devices in this article are for low-dimensional electron gas type systems such as 1D quantum wires and 2D quantum wells and heterostructures where confinement does not quantize electron dynamics in all three dimensions and there are still one (1D) or two (2D) translationally invariant dimensions producing a conserved wavevector. In ultrasmall quantum dot structures, on the other hand, confinement is three dimensional and electron dynamics is quantized in all three directions. Such quantum dot structures are better thought of as large artificial atoms or molecules whose many-body exchange-correlation effects are obtainable from quantum chemistry type configurational interaction calculations or from density functional calculations in three-dimensional confinement potentials. For quantum dot systems, the classical Coulomb energy is an important ingredient and the concept of Coulomb blockade seems to be widely applicable. For arrays of quantum dots, which form atomic or tight-binding bands, electron gas-type many-body theories discussed in this article become far too complicated and a model Hamiltonian approach based on the Mott-Hubbard model seems promising (Stafford and Das Sarma, 1994).

APPENDIX

We provide a very brief introduction to TDLDA response calculations (Marmorkos and Das Sarma, 1993; Ando, 1977; Ando and Mori, 1979; Katayama and Ando, 1985)

restricting our formalism to 2D quantum structures only (i.e., quantum wells, inversion layers, heterostructures, etc.). In the TDLDA theory the irreducible polarizability tensor is formally given by the ladder vertex correction formula (37), and, therefore, the dynamical screening function or the dielectric matrix has the formal structure (38):

$$\varepsilon = 1 - v\pi^0(1+\tilde{u}\pi^0)^{-1}$$

where, as before, v is the direct bare Coulomb interaction, $\tilde{u}$ is the exchange-correlation induced (short-range) vertex correction, and $\pi^0 \sim G^0G^0$ is the leading order irreducible polarizability. The key point is that all these matrices ε, v, π^0, $\tilde{u}$ are calculated in the LDA-density functional one electron basis by solving self-consistently the Schrödinger-like Kohn-Sham equations which contains in its potential energy operator the confinement potential (V_c), the self-consistent Hartree potential (V_H), and an LDA exchange-correlation potential (V_{xc}). To obtain the LDA basis for a 2D device one therefore solves the following set of equations self-consistently (we take z direction to be along the confinement potential with the x-y plane being the plane of the 2D electron gas):

$$\left[-\frac{\hbar^2}{2m}\frac{d^2}{dz^2}+V_{eff}(z)\right]\xi_i(z) = E_i\xi_i(z) \quad , \tag{56}$$

$$V_{eff} = V_c + V_H + V_{xc} \quad , \tag{57}$$

$$\frac{d^2V_H(z)}{dz^2} = -\frac{4\pi e^2}{\varepsilon_0}\left[n(z)-n_I(z)\right] \quad , \tag{58}$$

$$n(z) = g\sum_i^0 N_i\left|x_i(z)\right|^2 \quad . \tag{59}$$

In (56)-(59), $\xi_i(z) = \langle z|i\rangle$ is the effective one-electron subband basis; the external confinement potential $V_c(z)$ is to be obtained from the details of the confinement problem (this by itself may be non-trivial for complicated structures); the self-consistent Hartree potential $V_H(z)$ is calculated by solving the Poisson's equation (58) where the electron density itself, $n(z)$, enters as the source term (note that ε_0 is the static background lattice dielectric constant) with $n_I(z)$ the positive charge density (dopants, distant gates, etc.) producing the free carriers (note that n_I must be there for charge neutrality); and, finally, V_{xc} is the all important exchange-correlation potential whose form needs to be specified to complete the self-consistent loop of (56)-(59). Much has been written (Jones and Gunnarsson, 1989; Williams and von Barth, 1983) on the density functional theory in general and on approximate forms for V_{xc} in particular. (It is, in fact, a vast subject with many books and review articles covering the topic!). In general, V_{xc} is a functional of the density $n(z)$ and, of course, the exact functional is unknown (because one must exactly solve the inhomogeneous many-body problem for the confined system to get the exact functional). In practice, the simplest approximation to V_{xc}, the so-called local-density-approximation (LDA), seems to work remarkably well. (The precise reasons for this impressive quantitative validity of LDA are not yet understood.) In LDA, one assumes that the functional V_{xc} is just a local function of $n(z)$, i.e., $V_{xc} \equiv V_{xc}[n(z)]$, and that this function can be evaluated from a corresponding many-body calculation for the homogeneous translationally invariant system at the same electron density. For our purpose, the approximation boils down to taking

$$V_{xc} \equiv \Sigma(k_F^0; E_F^*) \quad , \tag{60}$$

where Σ is the many-body self-energy of a homogeneous 3D electron gas (with the same bare effective mass m and lattice dielectric constant ε_0) with a constant density $n_0 \equiv n(z)$ with k_F^0 the 3D Fermi wavevector (for a density $n_0 \equiv n(z)$) and E_F^* the renormalized chemical potential (or the Fermi energy). Note that (60) is equivalent to stating that

$$V_{xc} \equiv E_F^* - E_F \quad , \tag{61}$$

i.e., V_{xc} is just the exchange-correlation correction to the bulk chemical potential. We emphasize that LDA is a rather crude approximation which seems to work extremely well. The exact form for V_{xc} is given by

$$V_{xc} \equiv \frac{\delta E_{xc}\{n(\mathbf{r})\}}{\delta n(\mathbf{r})} \quad , \tag{62}$$

where $E_{xc}\{n(\mathbf{r})\}$ is the (unknown) exact exchange-correlation total energy functional. Below we provide a particular parametrized form (Stern and Das Sarma, 1984) for V_{xc} which has been successful in applications to 2D ultrasmall devices:

$$V_{xc}(z) \equiv V_{xc}(n(z)) = -\left[1+0.7734x\ln(1+x^{-1})\right](2/\pi\alpha r_0) \quad , \tag{62}$$

where

$$\alpha = (4/9\pi)^{1/3}; x \equiv x(z) = r_0/21, \tag{63}$$

and

$$r_0 = r_s(z) = \left[\frac{4}{3}\pi a_B^3 n(z)\right]^{-1/3} \quad , \tag{64}$$

with $a_B = \varepsilon_0\hbar^2/2me^2$ as the effective Bohr radius and V_{xc} above is given in units of effective Rydberg with $Ry \equiv e^2/2a_B$.

Once V_{xc} is known, the basis set $\{i\}$ can be obtained by solving the self-consistent equations and one can then obtain the response function matrices to calculate dynamical screening, etc. In this LDA formalism, the short-range effective vertex correction $\tilde{u}$ is simply the local derivative (Marmorkos and Das Sarma, 1993) of the exchange-correlation potential:

$$\tilde{u} \equiv U_{xc} \equiv \frac{\partial V_{xc}}{\partial n(z)} \quad . \tag{65}$$

ACKNOWLEDGMENT

The author is extremely grateful to the United States Office of Naval Research (Drs. Larry Cooper and George Wright) and the United States Army Research Office (Dr. Mike Stroscio) for steadfast support of his research on ultrasmall semiconductor devices.

REFERENCES

Abrikosov, A. A., Gorkov, L. P., and Dzyaloshinski, I. E., 1963, "Methods of Quantum Field Theory in Statistical Physics," Prentice-Hall, New York.

Ando, T., 1977, *Z. Phys. B* 26:263.

Ando, T., and Mori, S., 1979, *J. Phys. Soc. Jpn.* 47:1518.

Ando, T., Fowler, A. B., and Stern, F., 1982, *Rev. Mod. Phys.* 54:437.

Das Sarma, S., 1991, in "Light Scattering in Semiconductor Structures and Superlattices," Ed. by D. J. Lockwood and J. F. Young, Plenum, New York, NATO ASI 273:499.

Das Sarma, S., and Vinter, B., 1981, *Phys. Rev. B* 24:549.

Das Sarma, S., and Vinter, B., 1982, *Phys. Rev. B* 26:960.

Das Sarma, S., and Xie, 1987, *Phys. Rev. B.* 35:9875.

Das Sarma, S., and Xie, 1988, *Phys. Rev. Lett.* 61:738.

Das Sarma, S., Jain, J. K., and Jalabert, R., 1988, *Phys. Rev. B* 37:4560.

Das Sarma, S., Jalabert, R., and Yang, S. R. E., 1990a, *Phys. Rev. B* 41:8288.

Das Sarma, S., Jain, J. K., and Jalabert, R., 1990b, *Phys. Rev. B* 41:3561.

Fetter, A. L., and Walecka, J. D., 1971, "Quantum Theory of Many-Particle Systems," McGraw-Hill, New York.

Haldane, F. D. M., 1981, *J. Phys. C* 14:2585.

Hedin, L., 1965, *Phys. Rev.* 139:A796.

Hedin, L., and Lundqvist, S., 1969, *in* "Solid State Physics," Vol. 23, Ed. by F. Seitz, D. Turnbull, and H. Ehrenreich, Academic, New York.

Hu, Y. K., and Das Sarma, S., 1992, *Phys. Rev. Lett.* 68:1750.

Hu, Y. K., and Das Sarma, S., 1994, *Phys. Rev. B* 48:5469.

Hu, Y. K., and Das Sarma, S., 1994, *Phys. Rev. B* 48:14388.

Jain, J. K., and Das Sarma, S., 1987, *Phys. Rev. B* 36:5949.

Jain, J. K., and Das Sarma, S., 1987, *Surf. Sci.* 196:466.

Jalabert, R., and Das Sarma, S., 1989, *Phys. Rev. B* 40:9723.

Jones, R. O., and Gunnarson, O., 1989, *Rev. Mod. Phys.* 61:689.

Kadanoff, L. P., and Baym, G., 1962, "Quantum Statistical Mechanics," Benjamin, New York.

Katayama, S., and Ando, T., 1985, *J. Phys. Soc. Jpn.* 54:1615.

Li, Q. P., and Das Sarma, S., 1991, *Phys. Rev. B* 43:11768.

Lindhard, J., 1954, *K. Dan. Vidensk. Selsk. Mat. Fys. Medd.* 28:8.

Mahan, G. D., 1981, "Many-Particle Physics," Plenum, New York.

Marmorkos, I. K., and Das Sarma, S., 1991, *Phys. Rev. B* 44:3451.

Marmorkos, I. K., and Das Sarma, S., 1993, *Phys. Rev. B* 48:1544.

Pine, D., and Nozieres, P., 1966, "The Theory of Quantum Liquids," Benjamin, New York.

Quinn, J. J., and Ferrell, R. A., 1958, *Phys. Rev.* 112:812.

Rice, T. M., 1965, *Ann. Phys. (N. Y.)* 31:100.

Schrieffer, J. R., 1964, "Theory of Superconductivity," Addison-Wesley, New York.

Sólyom, J., 1979, *Adv. Phys.* 28:201.

Stafford, C. A., and Das Sarma, S., 1994, *Phys. Rev. Lett.* 72:3590.

Stern, F., 1967, *Phys. Rev. Lett.* 18:546.

Stern, F., and Das Sarma, S., 1984, *Phys. Rev. B* 30:840.

Vinter, B., 1976, *Phys. Rev. B* 13:4447.

Vinter, B., 1977, *Phys. Rev. B* 15:3947.

Williams, A. R., and von Barth, V., 1983, *in* "Theory of Inhomogeneous Electron Gas," Ed. by S. Lundqvist and N. H. March, Plenum, New York.

EFFECTS OF BAND-STRUCTURE AND ELECTRIC FIELDS ON RESONANT TUNNELING DYNAMICS

Jun He[1] and Gerald J. Iafrate[2]

[1]Department of Electrical and Computer Engineering
North Carolina State University
NC 27695-7911, U.S.A.
[2]U.S. Army Research Office
Research Triangle Park
NC 27709-2211, U.S.A.

I. INTRODUCTION

Tunneling phenomena in quantum wells and superlattices have been studied extensively for potential applications in quantum devices. Due to the many contemporary resonant tunneling experiments performed on double barrier structures, many analytical analyses and numerical simulations have been focused on such structures [1-3]. Theoretical calculations have affirmed that resonant tunneling occurs for the free electron in the double barrier structure [1], and for the Bloch electron in the one-dimensional nearest neighbor tight binding band with double potential barriers [2].

In this paper, we present a theory for describing Bloch electron transport in inhomogeneous electric fields due to localized impurities, and the dynamics of the Bloch electron in homogeneous electric fields in the presence of such impurities. Specifically, we consider the tunneling of a Bloch electron through a single and a double Slater-Koster type impurity potential barriers respectively [4], and tunneling of a Bloch electron in the above mentioned structures under the influence of external electric fields. Usually, the Bloch representation is used for the calculation of quantum transport involving the Bloch electron in a homogeneous electric field. However, it is difficult to use the Bloch representation in situations where the inhomogeneity is localized and non-perturbative in strength, such as localized impurities. Iafrate *et. al.* [5] have derived a novel theory for treating Bloch electron dynamics and quantum transport in inhomogeneous electric fields of arbitrary strength and time dependence, which includes all possible quantum effects, *i. e.*, intraband and interband scattering, interband Zener

tunneling, and non-linear transient transport. In this paper, we have apply the formalism previously developed by Iafrate *et. al.* [5]. In the formalism, the electric field is described through the use of the vector potential. The choice of the vector potential gauge leads to a natural set of basis function for describing Bloch electron dynamics. The formalism uses the localized Wannier representation, in which, the complete set of Wannier functions is used as basis function. These functions are inherently localized, thus making them most convenient for discussing localized inhomogeneous fields and the localized states arising from the inhomogeneity.

In the localized Wannier representation, the wave function of the electron is expressed as

$$\psi(\vec{r},t) = \sum_n \sum_{\vec{l}} f_n(\vec{l},t) W_n(\vec{r}-\vec{l},t) \ , \tag{1}$$

where $f_n(\vec{l},t)$ is the envelope function in the instantaneous Wannier representation, $W_n(\vec{r}-\vec{l},t)$ is the time-dependent Wannier function, and "n" indexes the band. The general differential equation for the time-dependent envelope function was derived in ref.[5], and in the single-band approximation, the differential equation is given by

$$i\hbar \frac{\partial f_n(\vec{r},t)}{\partial t} = \epsilon_n(-i\vec{\nabla} - \frac{e}{\hbar c}\vec{A}) f_n(\vec{r},t) + \sum_{\vec{l}'} eV_n(\vec{l}',\vec{r},t) f_n(\vec{l}',t) \ , \tag{2}$$

where $\epsilon_n(\vec{K})$ is the Bloch energy band function with crystal momentum $\vec{K}$, $V_n(\vec{r}',\vec{r},t)$ are the matrix elements of the inhomogeneous potential with respect to the localized basis defined by

$$V_n(\vec{l},\vec{l}',t) = \int d\vec{x} W_n^*(\vec{x}-\vec{l}',t) V(\vec{x},t) W_n(\vec{x}-\vec{l},t) \ ; \tag{3}$$

$\vec{A}$ is the vector potential due to the field of the spatially homogeneous part, $\vec{E}$, with $\vec{A} = -c\int_0^t \vec{E} dt'$.

To illustrate the physics of a Bloch electron tunneling through potential barriers under the influence of an external homogeneous electric field, we have applied the formalism to the single and the double Slater-Koster type impurity potentials in the one-dimensional lattice structure within the single-band nearest-neighbor tight-binding approximation; although we choose the Slater-Koster potential for simplicity here, the method can be extended to general localized inhomogeneous potentials. We derived the Green's function for the zero-field ($\vec{E}=0$) case, as well as the field-dependent Green's function for the constant field ($\vec{E} = \vec{E}_o$) case within the one-dimensional nearest-neighbor tight-binding approximation. Using the Green's function, we constructed the envelope function and calculated the transmission coefficients. The exact transmission coefficients for the single and the double impurity barriers are obtained for the zero-field nearest-neighbor tight-binding band, showing the resonances for the double barrier structure; the Stark energy spectra of Bloch electrons in electric fields with impurities are calculated for the single and the double Slater-Koster impurity structures. The time evolution of the envelope function due to the impurities as well as the external fields are examined. The results for the zero-field and the constant field calculations are presented in section II and section III respectively. In section IV, we employ previously reported results which identified band deformation caused by a superimposed DC and AC electric field [6]; we also study the DC component of the current derived therefrom, and establish the requisite transmission coefficients for tunneling through the impurity

barriers under specific tuning condition. In section V, we give a brief summary and concluding discussion.

II. TUNNELING OF THE BLOCH ELECTRON IN ZERO ELECTRIC FIELD

With the electric field turned off, the tunneling or scattering of a Bloch electron due the localized inhomogeneous potentials in the given energy band is a stationary problem. The transmission coefficients and the current densities for the Bloch electron tunneling through the potential barriers can be calculated through the use of the Green's function method. In this section, we will outline the method for examining the transport of a Bloch electron in the one-dimensional nearest-neighbor tight-binding band with single and double impurity potential barriers. It is noted that for $\vec{E} = 0$, the vector potential in eq. (2) $\vec{A} = 0$, and the time-dependence of eq. (2) is separable. Seeking a solution for eq. (2) in the form

$$f_n(\vec{r}, t) = F_n(\vec{r}) e^{-\frac{i}{\hbar} E_n t} \ , \tag{4}$$

where $F_n(\vec{r})$ is the time-independent part of the envelope function, and E_n is the separation constant (the energy of the Bloch electron), we obtain a time-independent differential equation for $F_n(\vec{r})$,

$$(E_n - \epsilon_n(-i\vec{\nabla}))F_n(\vec{r}) = \sum_{\vec{r}'} V_n(\vec{r}', \vec{r}) F_n(\vec{r}') \ , \tag{5}$$

where the sum on the right-hand side covers the entire range of the inhomogeneous potential. This equation can be used to treat any localized inhomogeneous potential. Ideally, if $V_n(\vec{r}', \vec{r})$ is confined to a few lattice spacing, we will have a finite difference problem of minor degree of difficulty. We may solve this equation for the impurity problems by finding the Green's function for the unperturbed Hamiltonian (the system without the impurities), writing the solution of eq. (5) in the Lippmann-Schwinger form, and determining the impurity levels and Bloch electron envelope function from the Lippmann-Schwinger equations.

To find the Green's function $G_o(\vec{r}, \vec{r}')$ for the unperturbed Hamiltonian of eq. (5), we look for the $G_o(\vec{r}, \vec{r}')$ which satisfies the equation

$$[\epsilon_n(-i\vec{\nabla}) - E_n] G_o(\vec{r}, \vec{r}') = -\delta_{\vec{r},\vec{r}'} \ , \tag{6}$$

for the given energy band $\epsilon_n(\vec{K})$. It can be shown that the solution to this inhomogeneous differential equation is,

$$G_o(\vec{r}, \vec{r}') = \frac{1}{N} \sum_{\vec{K}} \frac{e^{i\vec{K}\cdot(\vec{r}-\vec{r}')}}{E_n - \epsilon_n(\vec{K})} \ , \tag{7}$$

where the summation is over all possible values of "$\vec{K}$".

To demonstrate the usefulness of the method, we will discuss the one dimensional linear chain model within the nearest-neighbor tight-binding approximation. The dispersion relation associated with the Bloch state in the one-dimensional nearest-neighbor tight-binding approximation is

$$\epsilon_n(K) = \epsilon_o + 2\epsilon_1 \cos Ka \ , \tag{8}$$

where "a" is the spacing between the atoms. By converting the sum in eq. (7) into an integral, the summation in eq. (7) can be evaluated for the energy dispersion of eq.(8) [4], and the explicit forms of Green's function are obtained. For energy values outside the unperturbed band, $\mid x \mid > 1$, where "x" defined by $x = (E_n - \epsilon_o)/2\epsilon_1$, the Green's function is found to be

$$G_o(r, r') = \frac{1}{2\epsilon_1} \frac{e^{-u|r-r'|}}{\sinh u} \ , \tag{9}$$

where "u" is defined by $\mid x \mid = \mid (E_n - \epsilon_o)/2\epsilon_1 \mid = \cosh u$, and the Green's function decays exponentially as the distance $\mid r - r' \mid$ increases. Whereas, for energies inside the unperturbed band, *i. e.*, $\mid x \mid < 1$, the retarded Green's function is

$$G_o(r, r') = \frac{i}{2\epsilon_1 \sin u} e^{iu|r-r'|} \ , \tag{10}$$

where "u" is defined by $x = (E_n - \epsilon_o)/2\epsilon_1 = \cos u$, and the Green's function refers propagation of the Bloch electron to "plus infinity". In the following, we will derive the solution of eq. (5) for the lattice with one and two Slater-Koster impurities. We write the solution for the time-independent envelope function in the Lippmann-Schwinger form as,

$$F_n(r) = g F_n^o(r) + \sum_l \sum_{l'} V_n(l', l) G_o(r, l) F_n(l') \ , \tag{11}$$

where $F_n^o(r) = e^{iur}$, is the time-independent envelope function for the unperturbed Bloch state in the Wannier representation, the summations $\sum_l$, $\sum_{l'}$ are over all lattice sites. "g" is a constant, for in-band energy, g=1, corresponding to continuum states; for in-gap energy, g=0, corresponding to discrete bounded states.

First, as an example, we consider the well-known Slater-Koster localized potential [4], *i.e.*,

$$V_n(r', r'') = V_o \delta_{r',l_o} \delta_{r'',l_o} \ , \tag{12}$$

which indicates a single impurity of strength V_o located on the site l_o. For the in-gap energy, the time-independent envelope function for Bloch electron at any lattice site "r" is

$$F_n(r) = V_o G_o(r.l_o) F_n(l_o) \ . \tag{13}$$

Letting $l = l_o$ in eq.(13), and using eq.(9) for $G_o(r, l_o)$, we obtained the impurity energy levels for the single Slater-Koster impurity as,

$$E_n = \epsilon_o \pm 2\epsilon_1 \sqrt{1 + (\frac{V_o}{2\epsilon_1})^2} \ . \tag{14}$$

The envelope function for the bound state decays as the distance from the impurity increases,

$$F_n(r) = e^{-u|r-l_o|} F_n(l_o) \ , \quad u = \cosh^{-1} \sqrt{1 + (V_o/2\epsilon_1)^2} \ ; \tag{15}$$

with a decay length of $\lambda \sim 1/\cosh^{-1}\sqrt{1+(V_o/2\varepsilon_1)^2}$.

However, if the energy lies within the band, the envelope function can be written as

$$F_n(r) = e^{iur} + V_o G_o^+(r,l_o)F_n(l_o) \ , \tag{16}$$

where e^{iur} is the solution of the homogeneous equation (eq. (5) without impurities). The matching condition at the site of impurity l_o, *i. e.*, the continuity of the envelope function at the impurity site l_o, leads to the solution of eq.(5),

$$F_n(r) = \begin{cases} e^{iur} + \sqrt{R}e^{-iur} = e^{iur} + \frac{i\frac{V_o}{2\varepsilon_1 \sin u}e^{i2ul_o}}{1-i\frac{V_o}{2\varepsilon_1 \sin u}}e^{-iur} & r < l_o \\ \sqrt{T}e^{iur} = \frac{1}{1-i\frac{V_o}{2\varepsilon_1 \sin u}}e^{iur} & r > l_o \end{cases} , \tag{17}$$

where R is the reflection coefficient for the Bloch electron reflected by the impurity barrier,

$$R = \frac{(V_o/2\varepsilon_1 \sin u)^2}{1+(V_o/2\varepsilon_1 \sin u)^2} \ , \tag{18}$$

and T is the transmission coefficient for the Bloch electron tunneling through the impurity,

$$T = \frac{1}{1+(V_o/2\varepsilon_1 \sin u)^2} \ . \tag{19}$$

It is noted that T is always less than one for the single impurity case, with no resonances for the single impurity barrier structure.

Unlike the single impurity case, some interesting phenomena arise when two Slater-Koster type impurities are present in the crystal, such as, impurity levels splitting for the outside-band energy eigenvalues, and resonant tunneling for specific in-band energies. The impurity potential for the two Slater-Koster impurities is

$$V(r',r'') = V_o\delta_{r',l_o}\delta_{r'',l_o} + V_1\delta_{r',l_1}\delta_{r'',l_1} \ , \tag{20}$$

where l_o, l_1 denote the locations of the impurities, and V_o, V_1 denote the strength of the impurities. In accordant with eq.(11), the solution of the time-independent envelope function for the in-gap energy is

$$F_n(r) = V_o G_o(r,l_o)F_n(l_o) + V_1 G_o(r,l_1)F_n(l_1) \ . \tag{21}$$

The continuity of the envelope function at lattice sites l_o and l_1 give rise to two linear equations for $F_n(l_o)$ and $F_n(l_1)$, and the non-trivial solutions of $F_n(l_o)$ and $F_n(l_1)$ require the determinant of the coefficient matrix of the linear equations be zero,

$$\begin{vmatrix} 1-V_oG_o(l_o,l_o) & -V_1G_o(l_o,l_1) \\ -V_oG_o(l_1,l_o) & 1-V_1G_o(l_1,l_1) \end{vmatrix} = 0 \ . \tag{22}$$

There are four roots associated with the above equation which correspond to the impurity energy levels. If $V_1 = V_o$, the impurity energy levels can be expressed in a more explicit form as,

$$\begin{aligned} E_1^{\pm} &= \varepsilon_o + 2\varepsilon_1[1+(\frac{V_o}{2\varepsilon_1})^2(1\pm e^{-ul})^2]^{1/2} \ , \\ E_2^{\pm} &= \varepsilon_o - 2\varepsilon_1[1+(\frac{V_o}{2\varepsilon_1})^2(1\pm e^{-ul})^2]^{1/2} \ ; \end{aligned} \tag{23}$$

where "l" is the distance between the two impurities. Since $u = \cosh^{-1}|(E_n^{\pm}-\varepsilon_o)/2\varepsilon_1|$, eq. (23) is a transcendental equation of $E_{1(2)}^{\pm}$, and $E_{1(2)}^{\pm}$ must be solved numerically. It can be shown that the two energy levels $E_{1(2)}^{+}$ correspond to the symmetric functions of $F_n(r)$ defined about the center of these impurities, and $E_{1(2)}^{-}$ are corresponding to the anti-symmetric functions of $F_n(r)$ about the center of these two impurities.

For the energy E_n within the band $\epsilon_n(K)$, *i.e.*, the energies belong to a continuous energy spectrum, we therefore consider the transmission and the reflection of the Bloch electron by the impurity barriers. The envelope function for this case consists of two parts,

$$F_n(r) = e^{iur} + [V_oG_o(r,l_o)F_n(l_o) + V_1G_o(r,l_1)f_n(l_1)] \ , \tag{24}$$

where the first part e^{iur} is the solution for the unperturbed Hamiltonian, and the second part is due to the scattering from the inhomogeneities. Using the continuity conditions of the envelope function at sites l_o and l_1, we obtain the solution of eq.(5) for the Bloch electron with in-band energy for the double impurity barrier structure. The time-independent envelope function is found to be

$$F_n(r) = \begin{cases} e^{iur} + \frac{e^{i2ul_o}[i\alpha(1-i\beta)+i\beta(1+i\alpha)e^{i2ul}}{(1-i\alpha)(1-i\beta)+\alpha\beta e^{i2ul}}e^{-iur} & r < l_o < l_1 \\ \frac{(1-i\beta)e^{iur}+[i\beta+\alpha\beta(1-e^{i2ul})]e^{-iur}}{(1-i\alpha)(1-i\beta)+\alpha\beta e^{i2ul}} & l_o < r < l_1 \\ \frac{1}{(1-i\alpha)(1-i\beta)+\alpha\beta e^{i2ul}}e^{iur} & r > l_1 \end{cases} , \tag{25}$$

where α and β are the notations introduced for simplicity,

$$\alpha = \frac{V_o}{2\varepsilon_1 \sin u} \ , \quad \beta = \frac{V_1}{2\varepsilon_1 \sin u} \ . \tag{26}$$

The transmission coefficient for the Bloch electron tunneling through the two impurity barriers is

$$\begin{aligned} T &= |\frac{1}{(1-i\alpha)(1-i\beta)+\alpha\beta e^{i2ul}}|^2 \\ &= \frac{1}{1+[2\alpha\beta\sin ul-(\alpha+\beta)\cos ul]^2+(\alpha-\beta)^2\sin^2 ul} \ . \end{aligned} \tag{27}$$

If $V_o = V_1$, the two impurities are with the same strength,

$$T = \frac{1}{1+4(\frac{V_o}{2\varepsilon_1})^2[\frac{V_o}{2\varepsilon_1 \sin u}\sin ul-\cos ul]^2} \ . \tag{28}$$

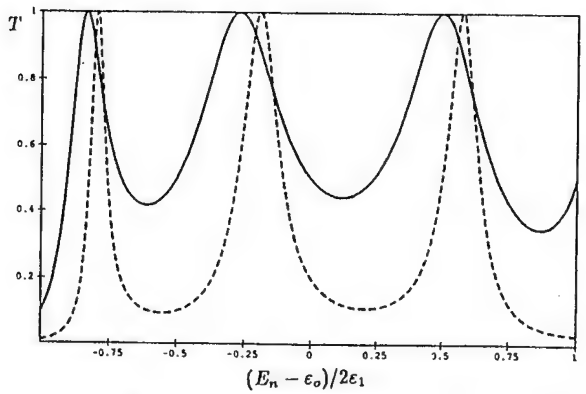


Figure 1. Transmission coefficients versus normalized energy for Bloch electron tunneling through double impurity barriers. The distance between the two impurities is $l = 4$ lattice spacing; the relative barrier height is $V_o/2\varepsilon_1 = 0.5$ (solid line), and $V_o/2\varepsilon_1 = 1.0$ (dashed line).

The resonant transmission (T=1) occurs when the energy of the electron E_n satisfies condition

$$\frac{V_o}{2\varepsilon_1 \sin u} \sin ul - \cos ul = 0 \ . \tag{29}$$

In fig. 1, we plotted the transmission coefficients T as a function of the normalized energy $(E_n - \varepsilon_o)/2\varepsilon_1$ of the Bloch electron for the double impurity structure with the spacing between the two impurities $l = 4$. It has been observed from our calculations that the number of resonance for the full range of the in-band energies equals the number of atoms between the two impurities, *i.e.*, "$l - 1$"; and the width of the resonances decreases as the impurity strength is increased.

In this section, we have demonstrated a method for treating the localized inhomogeneous potential through detailed study of one and two Slater-Koster type impurities in the one-dimensional linear chain lattice. It is very interesting to note that the tunneling of a Bloch electron through the impurity barriers is very analogous to the tunneling of a free electron through δ-function barrier. We can also show that the average velocity of the Bloch electron expressed in terms of the envelope function in Wannier representation for the nearest-neighbor tight-binding band is,

$$< v_n >= -\frac{2\varepsilon_1 a}{\hbar} Im\{\sum_l F_n(l) F_n^*(l-1)\} \ . \tag{30}$$

We have calculated the average velocities for the single and the double impurity barriers structures following eq. (30), and found that the average velocity of the Bloch electron equals the product of the transmission coefficient and the average velocity of the Bloch electron for the structure without impurity, *i.e.*, $< v_n >= T < v_n^o >$, where $< v_n^o >= -\frac{2\varepsilon_1 a}{\hbar} \sin u$ is the average velocity for the Bloch electron in the lattice without impurities. In addition to the single and the double impurity structures, we have explicitly derived the transmission coefficient for "L" contiguous impurities of the same

strength distributed along the lattice sites (double hetero-junction like structure) and an infinite number of impurities distributed along the half space (hetero-junction like structure) using the same method; the results will be published elsewhere.

III. EFFECTS OF THE DC FIELDS

When a constant homogeneous electric field E_o is applied to the crystal, the vector potential is linearly time dependent $A = -cE_o t$. This constant field causes Bloch oscillations for the system in the single band approximation while the energy spectrum forms a Wannier-Stark ladder with the spacing between ladders being "$eE_o a$". The differential equation for the time-dependent envelope function in the Wannier representation in the single-band model is [5],

$$i\hbar \frac{\partial f_n(l,t)}{\partial t} = \sum_{l'} [\varepsilon_n(l'-l,t) + V_n(l,l',t)] f_n(l',t) \ , \tag{31}$$

where $\varepsilon_n(l-l',t)$ is the Fourier component of the time-dependent energy band function $\epsilon_n(K - \frac{e}{\hbar c}A)$. Note that

$$\varepsilon_n(l'-l,t) = e^{i\omega_B(l'-l)t} \varepsilon_n(l'-l,0) \ , \quad V_n(l,l',t) = e^{i\omega_B(l'-l)t} V_n^o(l,l',0) \ , \tag{32}$$

where $\varepsilon_n(l'-l,0)$, $V_n^o(l,l',0)$ are time-independent, and ω_B is the Bloch frequency $\omega_B = eE_o a/\hbar$. We may separate the time-dependence of eq.(31) by writing the envelope function in the form of

$$f_n(l,t) = F_n(l) e^{-i(\frac{\varepsilon}{\hbar} + l\omega_B)t} \ , \tag{33}$$

where "ε" is a constant of separation, and the time-independent part of the envelope function $F_n(l)$ satisfies difference equation

$$(\varepsilon + l\hbar\omega_B) F_n(l) = \sum_{l'} [\varepsilon_n(l'-l) + eV_n^o(l,l')] F_n(l') \ . \tag{34}$$

In essence, eq. (34) depicts an infinite set of equations for $\{F_n(l)\}$, which is mathematically solvable for the special case of the nearest-neighbor tight-binding approximation. For the nearest-neighbor tight-binding band with the band dispersion given by eq. (8), eq.(34) is reduced to

$$(\varepsilon + l\hbar\omega_B) F_n(l) = \varepsilon_o F_n(l) + \varepsilon_1 F_n(l+1) + \varepsilon_1 F_n(l-1) + \sum_{l'} V_n^o(l,l') F_n(l') \ , \tag{35}$$

which can be solved by the Green's function method [9]. To bulid the Green's function for eq. (35), we need to find the solution of the homogeneous equation

$$(\varepsilon + l\hbar\omega_B) F_n^{(o)}(l) = \varepsilon_o F_n^{(o)}(l) + \varepsilon_1 F_n^{(o)}(l+1) + \varepsilon_1 F_n^{(o)}(l-1) \ . \tag{36}$$

Since the recurrence relations of $F_n^{(o)}(l)$ in equation (36) are the same as that of the Bessel functions [7], the general solution of eq.(36) is any linear combination of the Bessel functions of the first and the second kind, with index "$l + y$", where $y = \frac{\varepsilon - \varepsilon_o}{\hbar\omega_B}$;

and argument "$2\varepsilon_1/\hbar\omega_B$". For an infinite crystal, the boundary condition that F_n not diverge for both limit $l \to \pm\infty$ requires $y = \{m\}$, m=integer [8]. Therefore $J_{l+m}(\frac{2\varepsilon_1}{\hbar\omega_B})$ is the eigenfunction of equation (36) with an eigenvalue $\varepsilon_m = \varepsilon_o + m\hbar\omega_B$. Since

$$\sum_m J_{l+m}J_{l'+m} = \delta_{l,l'} \ , \quad \sum_l J_{l+m}J_{l+m'} = \delta_{m,m'} \ , \tag{37}$$

the eigenfunctions $\{J_{l+m}\}$ form a complete set of orthognormal functions, which can be used as basis function for expansion of the envelope function and the Green's function [9]. Hence, the Green's function of the unperturbed system is,

$$G_o(l,l',\varepsilon) = \sum_m \frac{J_{l+m}J_{l'+m}}{\varepsilon - \varepsilon_m} = \sum_m \frac{J_{l+m}J_{l'+m}}{\varepsilon - \varepsilon_o - m\hbar\omega_B} \ . \tag{38}$$

The corresponding retarded (advanced) Green's function can be evaluated. Expressing "ε" in terms of "y", $y = (\varepsilon - \varepsilon_o)/\hbar\omega_B$; we obtained the retarded (advanced) Green's function,

$$G_o^{\pm}(l,l',y) = \frac{1}{\hbar\omega_B}\lim_{\delta\to 0}\sum_m \frac{J_{l+m}J_{l'+m}}{y - m \pm i\delta} \tag{39}$$

$$= \frac{\pi(-1)^{l_<}}{\hbar\omega_B} J_{-l_<-y}J_{l_>+y}\{P\frac{1}{\sin\pi y} \mp i\sum_m \delta(y-m)\} \ , \tag{40}$$

where $l_<$ ($l_>$) is the smaller (larger) one of l, l'. In deriving the result of eq. (40), we have made the Fourier transform of eq. (39), where the summation of eq. (39) have been evaluated, and the inverse transform resulted in the explicit expression of eq. (40). It is noted that $G_o^{\pm}(l,l',y)$ consists of a principal part and an imaginary part, the imaginary part of the Green's function is directly related to the density of the states as [9]

$$\rho(l,y) = \mp\frac{1}{\pi}Im\{G_o^{\pm}(l,l,y)\} = \frac{(-1)^{l_<}}{\hbar\omega_B} J_{-l_<-y}J_{l_>+y}\sum_m \delta(y-m) \ , \tag{41}$$

where the δ-functions in the density of the states indicates that the energy spectrum for the unperturbed system is Wannier Stark ladders. As the impurities are turned on, the general solution of eq. (35) is

$$F_n(l) = gF_n^{(o)}(l) + \sum_{l'}\sum_{l''} V_n^o(l',l'')G_o(l,l',y)F_n(l'') \ , \tag{42}$$

where $F_n^{(o)}$ is the solution of homogeneous equation (eq. (36)), g is constant. For $y \neq \{m\}$, g=0; for $y = \{m\}$, g=1. To calculate the discrete Wannier Stark levels for the system with single and double impurities, we set g=0 in eq. (42). For a single Slater-Koster impurity located at l_o,

$$F_n(l) = V_oG_o(l,l_o,y)F_n(l_o) \ , \tag{43}$$

the continuity condition of $F_n(l)$ at $l = l_o$ requires

$$1 - V_oG_o(l_o,l_o,y) = 0 \ , \tag{44}$$

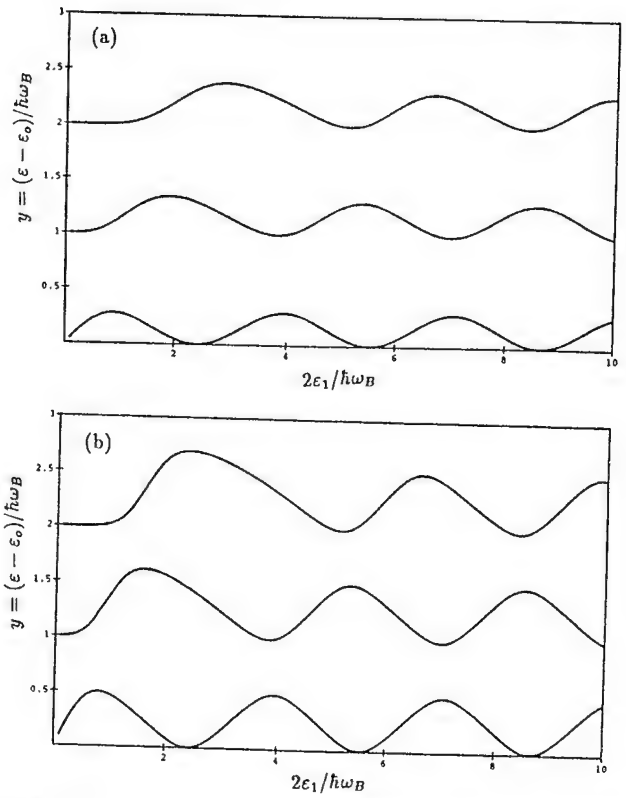


Figure 2. Energy spectra $y = (\varepsilon - \varepsilon_o)/\hbar\omega_B$ versus $2\varepsilon_1/\hbar\omega_B$ for the single impurity barrier, the relative impurity strength is (a) $V_o/2\varepsilon_1 = 0.5$, (b)$V_o/2\varepsilon_1 = 1.0$.

where "y", the energy eigenvalues for the single impurity structure, can be solved numerically. Results of solving eq.(44) can be found in fig. 2. M. Luban and J. H. Luscombe derived the similar form as that of eq. (44) for calculating the energy eigenvalues of a single impurity system by a different approach [10].

For two Slater-Koster impurities located at l_o and l_1,

$$F_n(l) = V_oG_o(l,l_o,y)F_n(l_o) + V_1G_o(l,l_1,y)F_n(l_1) \ , \tag{45}$$

the continuity of $F_n(l)$ at $l = l_o$ and $l = l_1$ results in two linear equation of $F_n(l_o)$ and $F_n(l_1)$, and the non-trivial solutions of $F_n(l_o)$ and $F_n(l_1)$ require that

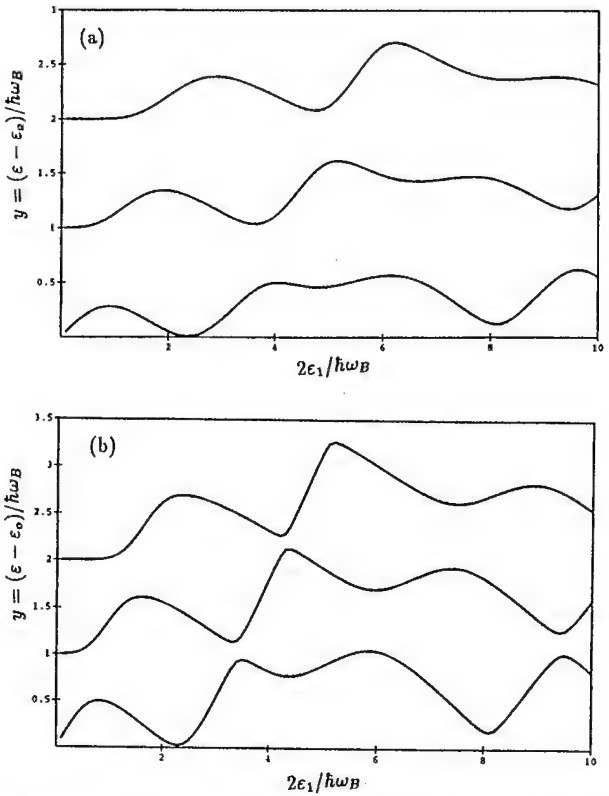


Figure 3. Energy spectra $y = (\varepsilon - \varepsilon_o)/\hbar\omega_B$ versus $2\varepsilon_1/\hbar\omega_B$ for the double impurities. The distance between the two impurities is $l = 4$ lattice spacing, and the relative impurity strength is (a) $V_o/2\varepsilon_1 = 0.5$, (b)$V_o/2\varepsilon_1 = 1.0$.

$$\begin{vmatrix} 1 - V_o G_o(l_o, l_o, y) & -V_1 G_o(l_o, l_1, y) \\ -V_o G_o(l_1, l_o, y) & 1 - V_1 G_o(l_1, l_1, y) \end{vmatrix} = 0 \ . \tag{46}$$

The energy spectrum for the double impurity barriers can also be solved numerically by finding roots of eq. (46). In fig. 3, we plot the energy spectrum "y" versus $2\varepsilon_1/\hbar\omega_B$ for the the double impurity structure with different impurity strengths. The results show that for relatively low potential barriers and high electric field, the energy level are almost uniform, resembling the Wannier-Stark ladder, and the barriers act as perturbations; for relatively low field and high potential barriers, the Wannier Stark uniformity

of the levels becomes distorted.

The non-uniformity of the Wannier Stark levels introduced by the presence of the impurities alters the time dependence of the envelope function for the Bloch electron. For the perfect crystal with no impurities, the time-dependent envelope function is

$$f_n^{(o)}(r,t) = \sum_{l'} K_n(r - l'; t, 0) f_n(l', 0) \ , \tag{47}$$

where $f_n(l', 0)$ is the initial value of the envelope function, determined by the initial condition of the Bloch electron, $K_n(l - l'; t, t')$ is the time evolution kernel for the Bloch electron in the unperturbed crystal. For the one-dimensional nearest-neighbor tight-binding band, the time evolution kernel can be expanded in the Bessel function basis (eigenfunction of the time-independent equation) as,

$$K_n(l - l'; t, t') = \sum_m J_{l-l'+m} J_m e^{-i(\frac{\varepsilon_o}{\hbar} + m\omega_B)(t-t') - i(l-l')\omega_B t'} \ . \tag{48}$$

So that the time-dependent envelope function $f_n^o(r,t)$ is a sum of simple harmonic oscillations with frequencies $\varepsilon_o/\hbar + m\omega_B$, which corresponds to the unperturbed Stark levels. When the impurities are present, the time-dependence of the envelope function will reflect the perturbations due to the impurities. The connection between the energy spectrum calculated for single and double impurity barrier (roots of eq. (44) and eq. (46)) and the time evolution of the Bloch electron becomes transparent when we look at the integral equation of the time-dependent envelope function,

$$\begin{aligned} f_n(r,t) &= \sum_{l'} K_n(r - l'; t, 0) f_n(l', 0) \\ &- \frac{i}{\hbar} \sum_{l'} \sum_{l} \int_0^t K_n(r - l; t, t') V_n(l, l', t') f_n(l', t') dt' \ . \end{aligned} \tag{49}$$

when $V_n(l, l', t)$ is a single or combination of Slater-Koster impurities, then $f_n(l', t')$ in the integrand, becomes the time-dependent envelope functions at the impurity sites, and can be evaluated through the use of the Laplace transforms. For the single and the double Slater-Koster impurity structures, we have derived the exact Laplace transforms of the envelope function at impurity sites, $L[f_n(l_o), s]$ and $L[f_n(l_1), s]$. To derive the explicit time-dependent envelope function at the sites of the impurities, one finds that the inverse Laplace transform of $L[f_n(l_o), s]$, $L[f_n(l_1), s]$ contain in the integrand an infinite number of poles along the imaginary axis in the complex plane of "s"; it is found that the locations of the poles coincide exactly with the energy spectra calculated from the distorted Wannier Stark ladder, *i.e.*, $s_m = i\varepsilon_o/\hbar + iy_m\omega_B$, y_m are the roots of eq. (44) (single impurity) or eq. (46) (double impurities). Using the residue theorem, one finds that the time-dependent envelope functions at the sites of the impurities have a time-dependence given by

$$f_n(l_o, t) = \sum_{m'} A_{m'}(l_o) e^{-i(\frac{\varepsilon_o}{\hbar} + y_{m'}\omega_B)t} \ , \tag{50}$$

$$f_n(l_1, t) = \sum_{m'} A_{m'}(l_1) e^{-i(\frac{\varepsilon_o}{\hbar} + y_{m'}\omega_B)t} \ ; \tag{51}$$

hence, the envelope functions at impurity sites are equal to summation of simple oscillations with frequencies $\varepsilon_o/\hbar + y_m\omega_B$. These frequencies mix into the time-dependent

envelope function at all other lattice sites through the integration over the envelope functions at impurity sites and the time-evolution kernel as in eq. (49). For example, for single Slater-Koster impurity,

$$f_n(r,t) = \sum_{l'} K_n(l-l';t,0) f_n(l',0) - \frac{i}{\hbar} V_o \int_0^t K_n(r-l_o;t,t') f_n(l_o,t') dt' \ , \tag{52}$$

where the integral $\int_0^t K_n(r-l_o;t,t') f_n(l_o,t') dt'$ mixes the frequencies of the Wannier Stark levels without impurities $\varepsilon_o/\hbar + m\omega_B$ and that with the impurity $\varepsilon_o/\hbar + y_m\omega_B$. The explicit time-dependence is obtained by substituting eq. (48) for $K_n(r-l_o;t,t')$ and eq. (50) for $f_n(l_o,t')$ into eq. (52), and evaluating the integral,

$$f_n(r,t) = \sum_{l'} K_n(l-l';t,0) f_n(l',0) - \frac{V_o}{\hbar\omega_B} \sum_m \sum_{m'} \frac{J_{r-l_o+m} J_m A_{m'}(l_o)}{m-r+l_o-y_{m'}} \left[e^{-i[\frac{\varepsilon_o}{\hbar}+(r-l_o+y_{m'})\omega_B]t} - e^{-i(\frac{\varepsilon_o}{\hbar}+m\omega_B)t} \right] \ , \tag{53}$$

where the $A_{m'}(l_o)$s are derived from the inverse Laplace transform of $L[f_n(l_o),s]$, and can be expressed as,

$$A_{m'}(l_o) = \frac{1}{\alpha_{m'}} \sum_{m''} \frac{C_{m''}(l_o)}{y_{m'}-m''} \ , \quad where\ \alpha_{m'} = -2\frac{V_o}{\hbar\omega_B} \frac{d}{dy} \left(\frac{\pi}{\sin \pi y} J_{-y} J_y \right)_{y=y_{m'}} \ ; \tag{54}$$

and $C_{m''}(l_o)$ are determined by the initial conditions as,

$$C_{m''}(l_o) = J_{m''} \sum_{l'} J_{l_o-l'+m''} f_n(l',0) \ . \tag{55}$$

IV. TRANSMISSION MODULATION BY AC FIELD TUNING

From our recent results on the theory of the Bloch electron dynamics in a superimpose uniform and oscillatory electric field [6], we are able to show that under special tuning conditions, a canonical band dispersion can be found for the band deformed by the superimposed electric field. This specialized tuning results in a DC component of the velocity for which we study the transport through single and double barriers. In this section, we will use the results to explore the transmission of the Bloch electron through single and double Slater-Koster impurity potential under the influence of the time-dependent electric field. The general form of the superimposed uniform and oscillatory electric field is

$$\vec{E} = \vec{E}_o + \vec{E}_1 \cos \omega t \ , \tag{56}$$

where E_o, E_1 are the strength of the DC and the AC field respectively, ω is the frequency of the AC field. In the following, we will again concentrate on the one-dimensional nearest-neighbor tight-binding band (eq. (8)). As the AC and the DC fields are tuned so that the Bloch frequency ω_B is a multiple of the AC frequency, *i.e.*, $\omega_B = M_o\omega$, where M_o is an integer, the equivalent band dispersion for the one-dimensional nearest-neighbor tight-binding band in the electric field is [6],

$$\tilde{\varepsilon}_n(K) = \varepsilon_o + 2\tilde{\varepsilon}_1 \cos Ka \ , \tag{57}$$

where $\tilde{\varepsilon}_1$ is the deformed band parameter, and

$$\tilde{\varepsilon}_1 = \varepsilon_1 (-1)^{M_o} J_{M_o}(\alpha) \ , \quad \alpha = \frac{eE_1 a}{\hbar\omega} \ . \tag{58}$$

It is clear that for the one-dimensional nearest-neighbor tight-binding band, the equivalent band dispersion $\tilde{\varepsilon}_n(K)$ has the same form as the original band dispersion, but with a different band parameter, or a modified band width; it is clear to see that the equivalent band width is $\tilde{W}_n = \mid J_{M_o}(\alpha) \mid W_n$. Since the absolute value of the Bessel function with integer order $\mid J_{M_o}(\alpha) \mid$ is always less than or equal to one, the applied electric field reduces the equivalent bandwidth.

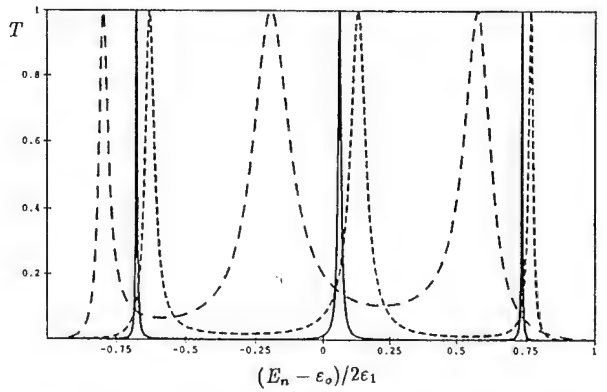


Figure 4. Transmission coefficient T versus the normalized energy $(E_n - \varepsilon_o)/2\varepsilon_1$ for the double impurity structure with the distance between the two impurities $l = 4$, and the relative impurity strength $V_o/2\varepsilon_1 = 1$. The longer-dashed line is for the zero field case, the small-dashed line is for $M_o = 1, \alpha = 2.0$, and the solid line is for $M_o = 1, \alpha = 0.5$.

Using the results of the zero-field transmission coefficients for single and double impurity barriers (eq. (19), eq. (28)) and the result of the band deformation (eq. (55)), we have obtained the transmission coefficients for the Bloch electron in the tuned electric fields. For single impurity, the transmission coefficient is

$$T = \frac{1}{1 + (\frac{V_o}{2\tilde{\varepsilon}\sin u})^2} \ ; \tag{59}$$

and for two Slater-Koster impurities, the transmission coefficient is

$$T = \frac{1}{1 + 4(\frac{V_o}{2\tilde{\varepsilon}_1 \sin u})^2 [\frac{V_o}{2\tilde{\varepsilon}_1 \sin u} \sin ul - \cos ul]^2} \quad . \tag{60}$$

It is noted that the modified band parameter always appears together with the impurity strength in transmission coefficients as $V_o/2\tilde{\varepsilon}_1$; and since

$$\frac{V_o}{2\tilde{\varepsilon}_1} = \frac{1}{(-1)^{M_o} J_{M_o}(\alpha)} \frac{V_o}{2\varepsilon_1} = \frac{\tilde{V}_o}{2\varepsilon_1} \quad , \quad \tilde{V}_o = \frac{V_o}{(-1)^{M_o} J_{M_o}(\alpha)} \quad , \tag{61}$$

then the deformed band parameter in the transmission coefficients can be treated as if the electron tunnels through the modified barriers in an unaltered band; in other words, the tuned electric field modifies the impurity barrier strength! As $\alpha = eE_1 a/\hbar\omega$ is tuned, $\mid J_{M_o}(\alpha) \mid$ will go through the values from zero to some maximum number (less than one), and the modified barrier strength $\tilde{V}_o$ will vary from a value greater than V_o to ∞. As a consequence, the transmission coefficients vary, and tend to have a narrowed width for the resonances. To illustrate the effects of the tuned AC and DC fields on the transmission coefficient, we take an example of $M_o = 1$, *i.e.*, $\omega_B = \omega$; by varying $\alpha = eE_1 a/\hbar\omega$ from 0 to 2, the value of $J_1(\alpha)$ varies from 0 to a maximum value of 0.5815 at $\alpha \approx 1.8$, and then decreases to 0.5767 at $\alpha \approx 2.0$. In fig. 4, we compare the transmission coefficients for $\alpha = 0.5$, $J_1(\alpha) = 0.2423$, $\alpha = 2.0$, $J_1(\alpha) = 0.5767$ with the zero field case. It is evident that the width of the resonances narrows significantly for the increased equivalent barrier height. It is noted that the minus sign of $(-1)^{M_o}$ has the effect of inverting the transmission coefficient about $(E_n - \varepsilon_o)/2\varepsilon_1$ for odd integer values of M_o.

This transmission modulation phenomena suggests the possibility of introducing three terminal effects in two terminal devices through the use of modulating fields; it also provides an example of the use of quantum control in affecting device characteristics.

V. DISCUSSION AND SUMMARY

In this paper, we have presented a theory for evaluating the transmission coefficient of a Bloch electron through impurity barriers. The zero-field Green's function is obtained for the one-dimensional nearest-neighbor tight-binding band, and the transmission coefficients are calculated for the single and the double Slater-Koster impurity barrier structures. The resonances are observed from the transmission coefficient for double barrier structure, the number of resonances equals the number of atoms between the two impurities, and the width of the resonances narrows for a increased impurity strength. In the theory, the envelope function in the Wannier representation is used to describe the Bloch electron; this description enable us to make good comparison between the in-band Bloch electron tunneling through the impurity barriers and the free electron tunneling through the δ-barriers.

The energy spectrum forms Wannier-Stark levels for the Bloch electron in a constant electric field, and the envelope function associated with each level is the first kind Bessel function with integer order for the infinite crystal, which serves as basis function for the envelope functions. The electric field-dependent Green's function is

derived within the one-dimensional nearest-neighbor tight-binding approximation, and Stark energy spectra are calculated for the one impurity and the two impurities cases. It is found that the time-dependent envelope functions at impurity sites equal sum of oscillations with frequencies corresponding to the energy spectra calculated; and the time-dependent envelope functions at all other sites equal sum of oscillations with frequencies corresponding to the unperturbed energy spectrum and the mixing terms which mix the energy spectrum with impurities and the energy spectrum without impurities.

The transmissions of the Bloch electron through the single and the double impurity barriers in a superimposed AC and DC field are analyzed. The band deformation due to the tuned electric fields, or equivalently, the change of the impurity barrier strength by the tuned fields, results in diminished transmission coefficient for single impurity barrier and narrowing of the width of the resonances in the transmission coefficient for the double impurity barriers.

ACKNOWLEDGMENT

This work was supported by Office of Naval Research and U. S. Army Research Office.

REFERENCES:

1. see, for example, E. Merzbacher, Quantum Mechanics, (Wiley, New York, 1970).
2. J. A. Støvneng and E. H. Hauge, *Phys. Rev.* B44, 13582 (1991)
3. W. R. Frensley, *Phys. Rev. Lett.* 57, 2853 (1986).
4. G. F. Koster and J. C. Slater, *Phys. Rev.*, 95, 1167 (1954).
5. G. J. Iafrate and J .B. Krieger, *Phys. Rev.* B40, 6144 (1989).
6. Jun He and Gerald J. Iafrate, *Bulletin APS* 39, No.1, R21 8, 894(1994), and to be published.
7. M. Abramowitz and I. A. Stegun, Handbook of Mathematical Functions, (Dover, New York, 1972).
8. H. Fukuyama, R. A. Bari, and H. C. Fogedby, *Phys. Rev.* B8, 5579 (1973).
9. E.N. Economou, Green's Function in Quantum Physics, 2nd Ed., (Springer, New York, 1990).
10. M. Luban and J. H. Luscombe, *Phys. Rev.* B34, 3674 (1986).

ARTIFICIAL IMPURITIES IN QUANTUM WIRES AND DOTS

A. S. Sachrajda[1], Y. Feng[1], G. Kirczenow[2], R. P. Taylor[1,3], B. L. Johnson[2], P. J. Kelly[1], P. Zawadzki[1] and P. T. Coleridge[1]

[1] Institute For Microstructural Sciences
National Research Council, Canada K1A OR6
[2] Department of Physics
Simon Fraser University, Canada V56 1S6
[3] School of Physics
University of New South Wales, Australia, NSW 2052

INTRODUCTION

One of the most common procedures employed in the fabrication of semiconductor nanostructures is the split-gate technique, first developed by Thornton *et al.* (1986). Submicron metallic gates are deposited on top of the semiconductor crystal [usually about 90 nm above the two dimensional electron gas (2DEG)] and are used to electrostatically define the nanostructure. Electrical contact is easily made to these 'top' gates in a region away from the nanostructure where they can be widened enough to accommodate the connecting wire. This technique allows important experimental parameters (such as potential barrier heights) to be controlled by varying the applied gate voltage. It has led to the observation of several novel effects including the quantization of the conductance of quantum point contacts (Wharam *et al.*, 1988; van Wees *et al.*, 1988). Recently we have added a new element to the split gate technology, i.e. the ability to contact 'isolated' submicron gates (or alternatively isolated ohmic contacts) (Feng *et al.*, 1993).

In this paper, the techniques necessary to fabricate and calibrate these structures are described. We demonstrate the versatility of this technique by reporting experiments in which isolated gates are used to create controllable potential variations inside a short quantum wire. Inhomogeneities of both types are studied, i.e. potential mounds (antidots) and depressions ('dimples'). In particular we concentrate on the quantum Hall regime in which the electrical transport is affected by magnetically bound states that exist in both the antidot and quantum dimple geometries. We observe several novel features which can be understood in terms of a theory (Sun and Kirczenow, 1994) based upon the tunnelling between edge and localized states.

FABRICATION AND CALIBRATION

Figure 1 is a schematic of the 'bridge' technique used to contact the isolated gate. The complete process involves the use electron-beam (e-beam) lithography, remote plasma enhanced chemical vapour deposition (RPECVD), reactive ion etching (RIE), and metallization process. The e-beam lithography and optical lithography are used to define features with widths less than and more than 5 microns respectively. Among the larger features defined by optical lithography are the gate bonding pads and Hall bars. Either polyimide or Si_3N_4 is used as the insulator for the bridge. The Si_3N_4 is deposited by RPECVD. And the RIE is used to open windows in the insulator.

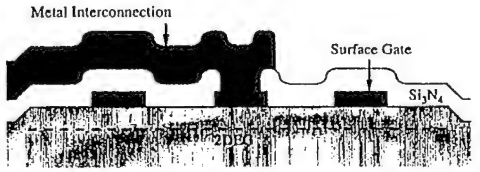


Figure 1. A schematic of the "bridge" technique used to contact isolated gates.

The fabrication begins with a high mobility GaAs/AlGaAs heterojunction grown by molecular beam epitaxy with a 2DEG typically 90 nm below the surface. The Hall bars and ohmic contacts are defined using conventional techniques. The top gates, including the isolated gate are written by e-beam lithography in polymethyl methacrylate (PMMA) bilayer resist using a converted JEOL JSM-840A SEM. At the same time alignment marks required for subsequent processing are defined (nb. the three e-beam stages used for this process require an alignment accuracy of 10 nm). This is followed by the deposition of 10 nm/40 nm of Ti/Pt and liftoff. The remaining processing is necessary to make a "bridge" contact to the isolated gate. First a layer with a thickness of about 100nm of Si_3N_4 (or polyimide in some devices) is deposited on the top of the surface gates as an insulator. By making the use of the highly accurate alignment markers a 100 nm diameter window is opened in the Si_3N_4 achieved by the second stage of e-beam lithography followed by RIE of the Si_3N_4. The third stage of ebeam lithography is then used to allow a deposition of the 100 nm wide metal link into the hole and thus establish the electrical connection between the center gate and the contact pad. The 100 nm thickness of Si_3N_4 was chosen so that voltages on this electrical link did not significantly affect the 2DEG underneath. For some applications which did not require a narrow link this third stage of e-beam could be replaced by conventional optical lithography. Recently this bridge technique has also been employed to make contact to isolated ohmic contacts in submicron 'Corbino' devices.

It is crucial to calibrate the potentials that one can control with the various gates. Such measurements are more complicated in multigate devices since in this case it is also important to identify how applying a voltage on one gate alters the potential between other sets of gates. The effect on the device resistance of depleting an individual submicron gate which only extends across half the sample (e.g. one gate of a split-gate device) is surprisingly large due to the effect of distorting the field lines around the gate. This allows one to determine the depletion voltages of all the 'individual' gates. The electron density in a constriction region formed between two gates can be obtained by sweeping the gates at magnetic fields that correspond to integer filling factors in the bulk. The plateaus that occur are simply related to the number of edge states (Haug *et al.*, 1988; Washburn *et al.*, 1988) (and hence the density) at that particular field. To determine the effect of applying a voltage

to a third gate on this calibration the measurements can be repeated with the third gate activated and the change in the plateaus positions noted. A variety of techniques exist to obtain the width of the constriction region. For simple point contact type of devices such as those described in this paper one can obtain the width by combining this calibration of density with the calibration of 1D subbands against gate voltage obtained at zero field (nb. an assumption needs to be made on the shape of the confining potential). Other calibration methods are specific to the geometry of the structure. One technique that we have made regular use of is a comparison of the low field magnetoconductance at temperatures around 4 K (where quantum interference effects are unimportant) with a semiclassical trajectory analysis (SCTA) based on the techniques of Baranger *et al.* (1991). We have used this method extensively to model quantum dots with a circular geometry (Taylor *et al.*, 1994) (or with a 'chaotic' elliptical geometry) and found that a simple central trajectory picture (i.e. one which only includes electron trajectories which enter the dot at 90° and assumes specular reflection) yields a good measure of the dot density. The full SCTA confirms this density and leads to a measure of the mean free path and beam collimation. We find that such a comparison is also useful for eliminating devices (Taylor *et al.*, 1993) which have intrinsic impurity potentials which randomize the electron trajectories resulting in a magnetoconductance that no longer reflects the geometry of the device e.g. a device that contains an impurity in a strategically important position such as close to the entrance or exit of the device. We are currently extending this type of analysis to model antidots and dimples. This is done by choosing a model potential for the interior which allows one to solve for the dynamics of the electron trajectory (Geisel *et al.*, 1992). Preliminary results have been obtained for an antidot structure where the interior is modeled by a Gaussian shaped hump in the center. By reversing the sign of the potential in the center of the wire we obtain a similar model for a quantum dimple.

MEASUREMENTS

In this section we present results from the antidot and dimple devices. Consider firstly the antidot system. When an antidot is placed within a nanostructure such as a quantum wire, then not only is a region of depleted electrons introduced but also two potential barriers are formed on either side between the antidot and the edges of the wire. At zero magnetic field these parallel conductors add approximately classically (i.e. the total conductance is the sum of the two individual conductances) in agreement with Simpson et al. (1993) [nb. there are predictions (Kirczenow, 1994) that under suitable experimental conditions the individual barriers cannot be considered separately and this simple behaviour will break down]. At low magnetic fields, the cyclotron diameter is larger than the diameter of the antidot. In this regime quantum interference effects are observed in the wire. The dominant period in the magnetoresistance, ΔB, is used to extract an area and hence diameter by means of the Aharonov-Bohm formula (i.e. $\Delta B = h/eA$). The diameter obtained in this way is found to be between that of the antidot and the wire width (Taylor *et al.*, 1994) confirming that the antidot plays the role of a single artificial impurity responsible for conductance fluctuations. This diameter can be made smaller or bigger by making the wire width smaller or the antidot diameter bigger respectively (Taylor *et al.*, 1994). At higher magnetic fields in the quantum Hall regime the electrical current in the Hall bar geometry is carried principally by the edge states (Büttiker, 1988). In very high magnetic fields in microscopic samples the coupling between different edge states at the same edge becomes negligible. This is the adiabatic regime of edge state transport. Consider the simple case, drawn schematically in Fig. 2(b) in which the voltages on the three gates are fixed so that only a single edge state is transmitted through each barrier. All other edge states are reflected at the barriers between the antidot and the wire. In the adiabatic regime since the mode around the antidot is not coupled to other current carrying modes the antidot plays no part in the conductance of the device other than, of course, being responsible for the presence of the barriers themselves.

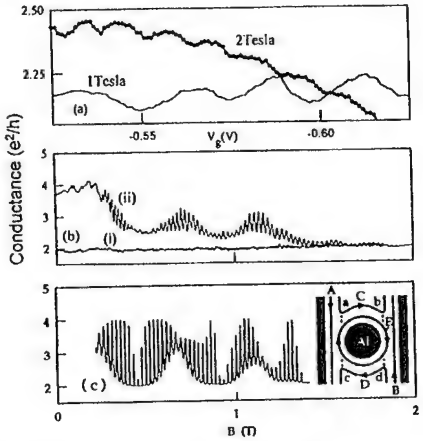


Figure 2. (a) Antidot voltage sweeps at 1 T and 2 T. (b)(ii) and (c) are the magnetoconductance experimental and theoretical traces respectively for the edge state configuration of 2(b). b(i) is the conductance when the constriction on one side of the antidot has been pinched off.

From the considerations above one would expect the conductance of the device to vary from a value of about $4e^2/h$ (the sum of the two $2e^2/h$ conductances) at zero and low magnetic fields to about $2e^2/h$ (the adiabatic value) at high fields. In the intermediate field regime the behaviour of the conductance is dependent on whether the coupling to the mode around the antidot is principally from those edge states which are reflected or the transmitted at the antidot. This would result in an increased or decreased total conductance respectively. Detailed experiments indicated (Kirczenow et al., 1994; Sachrajda et al., 1994) that in the intermediate field regime the antidot affects the conductance principally by increasing the transmission of the reflected edge states. We stress that although the transmission of edge states and hence the measured conductance is determined at the antidot, the location of the dissipation in such experiments occurs at the contacts. Figure 2(b)(ii) is a plot of the conductance of the antidot device under these conditions. The transition from around $4e^2/h$ at zero magnetic field to about $2e^2/h$ at high magnetic fields is clearly seen. Figure 2(c) is the result of a calculation (Sun and Kirczenow, 1994) in which scattering was assumed to take place at the high curvurture points points *a,b,c* and *d*. The theoretical analysis is based on a generalization of the edge-state scattering theories of Büttiker (1992) and Kirczenow and Castaño (1991). It is assumed that the current amplitude leaving any scattering event is related to the impinging current amplitude via a unitary scattering matrix. The current amplitude acquires a magnetic-field-dependent phase in transiting the path between scattering events. The current amplitude relationships as well as the unitary constraint on the scattering matrices, generate a set of equations which may be solved for the current amplitudes leaving the antidot in terms of the current amplitudes entering and the phase gained in circulating the various loops and the antidot. The high

frequency oscillations are due to antidot levels similar to the impurity levels predicted and observed previously in wires (Jain and Kivelson, 1988; Simmons *et al.*, 1989) and parallel constrictions (Smith *et al.*, 1989; Hwang *et al.*, 1991). They are also analogous to the zero dimensional states observed in quantum dots (van Wees *et al.*, 1989; Dharma-Wardana *et al.*, 1992).

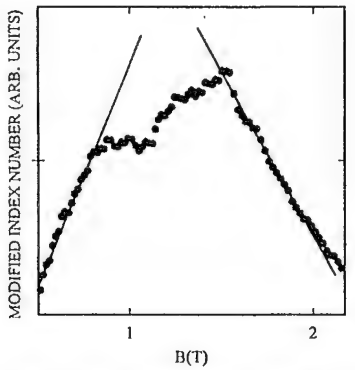


Figure 3. A modified index number plotted against the field at which the magnetoconductance resonances occur (see text for details). Three different slopes are visible. The change in slopes correspond to sharp increases in the magnetic field period of the oscillations. The solid lines are a guide to the eye.

Each resonance corresponds to a new level when the edge state around the impurity encloses an additional flux quantum. These can also be observed at fixed magnetic field by sweeping the size of the impurity. This is demonstrated in Fig. 2(a) which shows antidot gate voltage sweeps at two different magnetic fields. It is found experimentally that the gate voltage period is proportional to $1/B$. This can be understood by assuming that the depletion distance of the antidot gate varies linearly with gate voltage i.e. $\Delta r \propto \Delta V_g$ then for small changes in antidot radius, r, the change in area, ΔA, equals $2\pi r \Delta r \propto \Delta r$. Hence $\Delta V_g \propto \Delta A$ and since $\Delta A = h/eB$ for each resonance the period in gate voltage $\Delta V_g R \propto 1/B$. The beats in Figs. 2(b) and 2(c) are due to an inter-Landau level quantum interference effect (Kirczenow and Casataño, 1991). For example at point C, there is a loop due to scattering at points a and b. Such an effect was first observed in a four terminal quantum dot by Ford *et al.* (1991). Figure 3 illustrates a surprising effect which is not explained by the above model. The magnetic fields at which the individual conductance resonances (i.e. oscillation peaks) occurred were first plotted against a counting index number (i.e. the first peak was labelled 1, the second 2 etc...). A linear background was then removed and the residuals were then plotted to obtain the data in Fig. 3. Three distinct slopes can be seen in the plot. Each change in slope corresponds to a sudden increase in period (typically 15%). This change in period occurred at different magnetic fields on different cooldowns or after various amounts of illumination with a red light emitting diode. Although many simple explanations can account for decreases in period as a function of increasing magnetic field it is surprisingly difficult to explain a sudden increase in period. A more complete discussion is given elsewhere (Sachrajda *et al.*, 1994) but among the explanations which could not account for the data were; (a) a change in enclosed area due an inhomogeneous potential - this would always result in an increased area and therefore decreased period; (b) period changes associated with the beating behaviour (nb. such effects were automatically included in the calculations described above). (c) A $\Delta A \cdot B$ correction to the enclosed flux which occurs as a result of the magnetic field variation in the spatial position of the edge state and hence the area it encloses. An unlikely sharp feature in the antidot potential might therefore be expected to cause a period change - however calcuations showed that due to the finite edge state wavefunction width any sharp feature would result in a gradual change in period. It was found, however, that the above effect was consistent with a novel non-local effect which requires the presence of a second intrinsic potential inhomogeneity i.e. a partially depleted region is required. As the magnetic field is raised and the last Landau level begins to deplete in the region of the potential inhomogeneity an edge state starts to form around it. The presence of this new mode at the Fermi energy is found to have a significant effect on the rate at which the antidot modes pass through the Fermi energy (and hence on the period of the magnetoconductance oscillations) assuming (and in fact due to) a local conservation of charge (Kirczenow *et al.*, 1994).

At higher magnetic fields ~2 T we observed the beginning of spin splitting in the resonances. However, as the field was raised further the spacing of the resonances originating from the two spin species are found to lock exactly π out of phase. In addition they often lock in oscillation amplitude resulting in apparent period halved oscillations when compared to the low field data. This effect is similar to what we have previously observed and reported in quantum dots (van Wees *et al.*, 1989; Dharma-Wardana *et al.*, 1992). In that case the effect was explained qualitatively in terms of an inter-edge state Coulomb blockade effect (Sachrajda *et al.*, 1993). We speculate that a similar mechanism is responsible for the period halving effect in the present device. We note, as was pointed out by Simpson *et al.*(1993), that in this case this is a charging effect in an open geometry. At even higher magnetic fields the coupling between edge states becomes negligible and no further resonances are observed.

The above results involve an artificial inhomogeneity consisting of a submicron depleted region within the sample, i.e. an antidot. However, intrinsic inhomogeneities can also involve regions with an enhanced number of electrons. Using our bridging technique we can achieve and study an artificial impurity of this type by applying a positive voltage to the isolated gate. This we call a quantum 'dimple'. We have studied the effect of such a dimple in a potential barrier region as a function of the depth of the dimple potential. Different depths of potential are achieved by changing the magnitude of the positive gate voltage. We find that, as in the case of an antidot, a 'dimple' potential contains localized modes which affect the conductance. A full description of the results will be given elsewhere but we present below a brief summary and theoretical analysis for one potential depth. Figures 4(a), 4(b), and 4(c) contain an experimental magnetic field sweep, the results from a theoretical model and edge state schematics respectively, for one value of dimple gate voltage. The theoretical model is similar to that used for the antidot experiments. The theoretical results are expressed in terms of a dimensionless flux, the ratio of the flux threading the closed loop C to the flux quantum and is thus a function of both the area of the loop and the applied field. Consider the edge state configuration 4(c)(i). At low fields there are three edge states impinging on the dimple from each direction and a single localized state C. The edge states B and D are coupled to the localized state via scattering events 1,2 and 4. As the field is raised the state D depopulates. At still higher magnetic fields the edge state B begins to pinch off and state C depletes since the magnetic field pushes the allowed levels up in energy eventually pushing them through the Fermi energy. This leads to the situation shown in Fig. 4(c)(ii) where the edge state B now couples to the new localized state C' (which has developed out of B) through 2 and 5 while the edge state A couples to C' at the points of maximum curvature as shown. From here edge state B depopulates while at the same time the coupling between A and C' weakens and C' depopulates. In the theoretical analysis the

various couplings (e.g. the coupling of D and E to C) are varied smoothly to zero as the field is raised. We can now identify several regions of the experimental curves. The wide conductance minima in the low flux regime (an example is labelled "W" in the figures) are the result of interference between the closed loop states composed of state C and the possible loops including part of C and parts of B and F. For example the loop defined by the segments $C1$-$C6$-$F6$-$F3$-$C3$-$C1$ has a different length than the path around loop C. The phase difference accumulated in traversing the two paths can lead to constructive or destructive interference between the two paths. The coupling from B and F to C' in Fig. 4(c)(ii) vanishes more quickly than the others since the edge state B and F pinch off at lower fields than those which decouple state A and G from C'. The regular magnetoconductance oscillations at $P4$ are due to Aharanov-Bohm type interference associated with state C'. The drop in conductance from near $4e^2/h$ to near $2e^2/h$ at $P2$ is the result of the pinch off of modes B,F followed by the decoupling of A,G from mode C'. A slight minimum at the point marked 'M' has a similar origin to the 'W' minimum. The oscillatory structure on this minima is due to the persistent coupling of A,G to C' as mentioned above. This last feature is more apparent for deeper 'dimples' (Johnson *et al.*, 1994). The characteristics described above are for one particular dimple size. We show elsewhere that as the size of the dimple is changed the experimental features change in a way consistent with the above model (Johnson *et al.*, 1994).

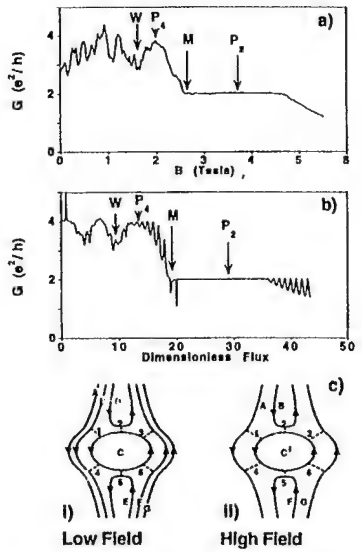


Figure 4. (a) A magnetoconductance plot for one value of gate voltage for the quantum "dimple" device. (b) The theoretical plot equivalent to (a) (see text). (c) edge state configurations in the device at two magnetic fields (see text).

In summary we have described the fabrication and calibration of devices which include isolated submicron gates. This technique is used to create local perturbations in the potential. We have presented results from two such experiments, involving an antidot and a 'dimple'. We have shown how both types of perturbation influence the magnetoconductance.

ACKNOWLEDGEMENTS

The authors thank M. W. C.Dharma-Wardana and D. Loss for stimulating discussions, and J. A. Adams, M. Davies, P. A. Marshall, P. Chow-Chong and R. Barber for assistance in device fabrication.

REFERENCES

Baranger, H. U., Vincenzo, D. P., Jalabert, R. A., and Stone, A. D., 1991, *Phys. Rev. B* 44:10637.
Büttiker, M., 1988, *Phys. Rev. B* 38:9375.
Büttiker, M., 1992, *in* "Semiconductors and Semimetals," Vol. 35, Academic Press, New York, 191.
Dharma-Wardana, M. W. C., Taylor, R. P., and Sachrajda, A. S., 1992, *Sol. State Commun.* 84:631.
Feng, Y., *et al.*, 1993, *Appl. Phys. Lett.* 63:1666.
Ford, C. J. B., *et al.*, 1991, *Phys. Rev. B* 43:7339.
Geisel, T., Ketzmerick, R., and Schedletzky, O., 1992, *Phys. Rev. Lett.* 69:1680.
Haug, R. J., MacDonald, A. H., Streda, P., and von Klitzing, K., 1988, *Phys. Rev. Lett.* 61:2797.
Hwang, S. W., *et al.*, 1991, *Phys. Rev. B* 44:13497.
Jain, J. K., and Kivelson, S. A., 1988, *Phys. Rev. Lett.* 60:1542.
Johnson, B. L., *et al.*, 1994, submitted for publication.
Kirczenow, G., 1994, *Phys. Rev. B*, in press.
Kirczenow, G., and Castaño, E., 1991, Phys. Rev. B 43:7343.
Kirczenow, G., *et al.*, 1994, *Phys. Rev. Lett.* 72:2069.
Sachrajda, A. S., *et al.*, 1993, *Phys. Rev. B* 47:6811.
Sachrajda, A. S., *et al.*, 1994, to be published.
Simmons, J. A., *et al.*, 1989, *Phys. Rev. Lett.* 63:1731.
Simpson, P. J., *et al.*, 1993, *Appl. Phys. Lett.* 63:3191.
Smith, C. G., *et al.*, 1989, *J. Phys. Cond. Matter* 1:6763.
Sun, Y., and Kirczenow, G., 1994, *Phys. Rev. Lett.* 72:2450.
Taylor, R. P., *et al.*, 1993, *Phys. Rev. B* 47:4458.
Taylor, R. P., *et al.*, 1994a, *Surf. Sci.* 305:648.
Taylor, R. P., Sachrajda, A. C., Freedman, D., and Kelly, P. J., 1994b, *Sol. State Commun.* 89:579.
Thornton, T. J., *et al.*, 1986, *Phys. Rev. Lett.* 56:1198.
van Wees, B. J., *et al.*, 1988, *Phys. Rev. Lett.* 60:848.
van Wees, B. J., *et al.*, 1989, *Phys. Rev. Lett.* 62:1181.
Washburn, S., Fowler, A. B., Schmid, H., and Kern, D., 1988, *Phys. Rev. Lett.* 61:2801.
Wharam, D. A., *et al.*, 1988, *J. Phys. C* 21:L209.

QUANTUM TRAFFIC THEORY OF SINGLE ELECTRON TRANSPORT IN NANOSTRUCTURES

John R. Barker and Sharif Babiker

Nanoelectronics Research Centre
Department of Electronics and Electrical Engineering
University of Glasgow
Glasgow G12 8QQ, Scotland, UK.

INTRODUCTION

Consider the possibility of a single electron tunneling through an insulating region between two metallic islands. For this to be possible the electron must somehow acquire a charging energy of the order of $e^2/2C$, where C is the effective inter-capacitance. At "high temperatures" where $T > T_c = e^2/2Ck_B$, the charging energy is easily supplied from thermal fluctuations. In the opposite extreme, the so-called Coulomb Blockade regime, $T << T_c$, the charging energy must be supplied externally for example by applying a voltage across the metal islands; otherwise the tunnelling is blocked. The simplest manifestation of the Coulomb blockade is thus a voltage offset $e/2C$ in the current-voltage characteristics. The most interesting effects however derive from considering the transport of more than one electron through an array of capacitors (Likhaerev, 1988). The transporting/tunnelling electrons become highly correlated in space and time since they must "queue up" to let a definite maximum number of electrons into the capacitor system at a time commensurate with the local charging-energy conditions. The motion is quite subtle because the polarisation field resulting from a propagating electron is distributed over all the metallic islands resulting in electrostatic soliton states forming in the electrode array. It is this solitonic property, a moving particle plus an associated field, which imparts a considerable stability on the electrical properties of a driven capacitor chain at temperatures $T << T_c$ (Barker et al., 1992a). It is not actually necessary for this effect to require tunnelling, any weak "conducting" path such as a charge leakage path, a hopping path or a bottleneck in the conductance map will suffice. It should also be noted that the condition $T << T_c$ must be supplemented by the condition that the series resistance must exceed the quantum resistance $R > R_Q = h/e^2$ for the effects of quantum fluctuations to be suppressed.

The controlled transport/correlated-tunnelling of single-electronic excitations in coupled tunnelling capacitator structures is now well-established experimentally in metal-insulator, metal-semiconductor systems and in capacitatively-coupled quantum point contacts in semiconductor 2DEG structures (Fulton and Dolan, 1987; Kuzmin *et al.*, 1989; Geerligs *et al.*, 1990; Meirev *et al.*, 1989). These structures rely on the existence of ultra-small capacitative structures such that the effective charging energies $e^2/2C$ exceed the thermal energy k_BT (Averin and Likhaerev, 1986). By exploiting state-of-the-art nanofabrication it is possible to construct 20 nm scale coupled capacitors (metal on semiconductor coupled Schottky dots) (Barker *et al.*, 1992b) which point the way to a future high temperature, high density nanoelectronic systems technology (Fig.1). From a theoretical point of view or from a modelling approach single electronic devices necessarily occur as inhomogeneous many-electron systems which makes their consideration problematic. Until recently the main modelling tool has been Monte Carlo simulation of the point tunnelling events coupled to electrostatic equations.

Recently we have introduced a new approach to modelling single-electronic systems that captures the Poisson stochastic nature of tunnel events and provides a fast, physically transparent and efficient method of calculating the steady-state characteristics of multi-junction configurations which in many cases allows exact solutions. This new method involves a re-formulation of the transport equations in terms of of *queuing theory* or *traffic theory*(Babiker and Barker, 1993) and centres on determining the distribution P_1 of quasi-electrostatic soliton excitations that are formed during the transport/tunnelling process.

SOLITON STATES IN SINGLE ELECTRONIC SYSTEMS

Consider an array of tunnel junctions represented by the simple multi-junction equivalent circuit shown in figure 3 . In the presence of excess charges the nodal potential equations are:

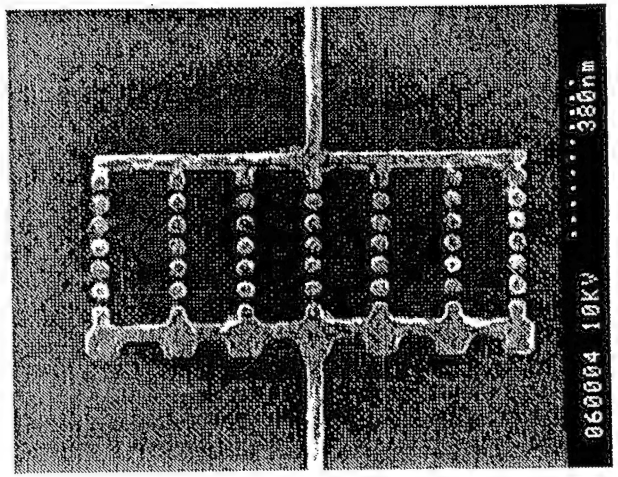
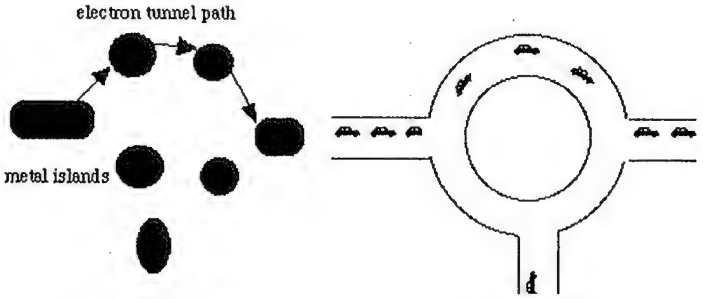


Figure 1. 40 nm diameter hemispherical Aluminium Schottky dots on p-Silicon with 15 nm spacing (Weaver, ???).

(a) double junction pass transistor (b) quadruple junction pass transistor

Figure 2. Schematic single-electronic device layout.



(a) Capacitor chains (b) Traffic flow on highways

Figure 3. Analogy between correlated tunnelling between conducting nodes and traffic theory.

$$\phi_1 = V_L \quad , \tag{1a}$$

$$C(\phi_i - \phi_{i-1}) + C_0\,\phi_i - C_i(\phi_{i+1} - \phi_i) = Q_i \tag{1b}$$

$$\phi_{N+1} = V_R \ . \tag{1c}$$

If a single excess charge (±e) exist at node k then from (1b)

$$C(\phi_{i+1} - 2\,\phi_i + \phi_{i-1}) - C_0\phi_i = (\pm e)\delta_{ik} \ , \tag{2}$$

which has the solution

$$\phi_i^{[k]} = \pm\frac{e}{2C\sinh(\lambda)}e^{-\lambda|i-k|} \ , \tag{3}$$

where $\lambda = \operatorname{arccosh}(1 + C_0/2C)$. Expression (3) represents a soliton with a length λ (units = number of junctions) with total charge ±e. A similar analysis may be made for more complex systems including two or three dimensional arrays by utilising the capacitance matrix.

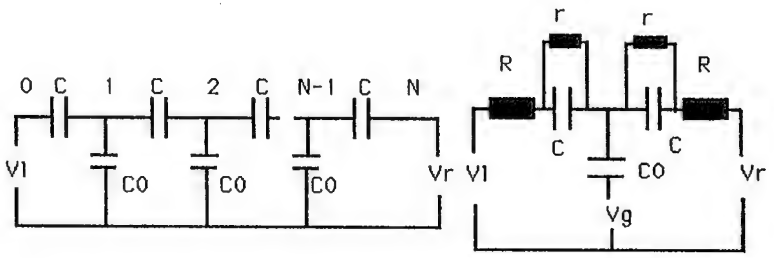


Figure 4. (a) A multi-junction tunnelling array. (b) Gated 2-junction array. The capacitances C are tunnel junctions. C_o is taken to be non-tunneling.

Expression (3) is essentially a Green function for the electrostatic equations. It follows that we may express the electrostatic state of the system by the potentials at each node: $\{\phi_i(t);\ i = 1...N\}$. From (3) we may write

$$\phi_i = \phi_i^{source} + \sum_k n_k \phi_i^{[k]} \ , \tag{4}$$

where $\phi_i^{source} = V_L\exp(-\lambda i)$ and where N is the number of nodes that can hold excess electrons (note that some nodes are tunnelling nodes, others are not) and n_k is the number of solitons present at node i; this number is an occupancy number and it may be positive or negative (anti-soliton). The n_i thus allow a description of the state of the system by occupation numbers

$$\Psi = \mathbf{n} = \{n_i(t);\ i=1...N\}. \tag{5}$$

TRAFFIC THEORY I - NODAL LEVEL

Let us now assume that after a tunnelling event the charge relaxes to a steady value in a short period $\tau_{relax} < \tau_{tunn}$ = time between tunnel events. We observe that during any period when the system is in a state $\Psi = \mathbf{n}$, electrons may tunnel from the ith node to the jth node provided i and j are connected by a tunnel path. We may define a connection matrix X_{ij} which assumes the values $X_{ij}=1$ if node i and node j are connected by a tunnelling path and otherwise is zero; by symmetry $X_{ij}=X_{ji}$.

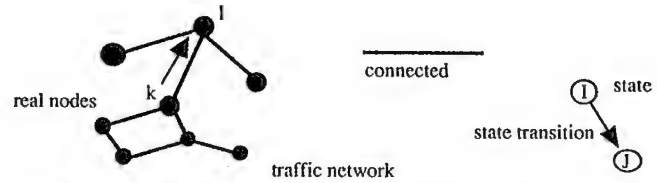


Figure 5. Tunnel events between real nodes and equivalent state transitions.

Suppose that the quantum mechanical transition probability per unit time for tunnelling from i to j in state $\mathbf{n}$ is $\Gamma(i,j;\mathbf{n})$, we may then define the probability that an electron is destined for node j by: $p_{ij}(n) = \Gamma(i,j;\mathbf{n})/\Gamma_i(\mathbf{n})$ where $\Gamma_i(\mathbf{n}) = \Sigma\, \Gamma(i,k;\mathbf{n})$ (sum over k: X_{ik} =1) gives the traffic throughput at the node i for a given soliton state $\mathbf{n}$. Similarly, incoming traffic to node i is governed by $\Lambda_i(\mathbf{n}) .= \Sigma\, \Gamma(k,i;\mathbf{n})$ (sum over k: X_{ik} =1). The analogy with conventional traffic theory (Kelly, 1991) is striking: the discrete charges move stochastically from one node to another where they queue up before being serviced (a further jump). In the quantum case the queues are always full and the departure process at a node is affected by the various service rates which depend on the state $\mathbf{n}$ of the system. Since the charges at a node are indistinguishable the quantum theory corresponds to the random selection queue discipline of classical traffic theory (Cooper, 1972; Kleinrock, 1976).For a constant $\mathbf{n}$, the arrival process is a Poisson process leading to a decaying exponential distribution of times between tunnel events. When the traffic intensity satisfies $a_i(\mathbf{n}) = \Lambda_i/\Gamma_i > 1$ electrons tend to accumulate at a node whereas for $a_i(\mathbf{n})< 1$ the tendency is for electron loss corresponding to positive charge accumulation.

The quantum traffic model is complicated by the fact that the state $\mathbf{n}$, and hence most of the array properties such as the $\Gamma(i,j;\mathbf{n})$, is generally changed by a tunnelling transition .In classical traffic theory of communications systems this corresponds to the broadcasting of the current state of a node to all other nodes and consequent updating of routing policy: adaptive routing theory (Kleinrock, 1976).

TRANSPORT PARAMETERS

The key parameter in quantum traffic theory is P_J the occupancy probability of the legal soliton state $\mathbf{n}_J$. All the usual transport properties may be derived from P_J, for example the current between two nodes l and m is given by

$$I_{lm} = e\sum_{J=1}^{M} P_J \Gamma(l,m;\mathbf{n}_J) \ , \tag{7}$$

and the average number of excess charges at node l is given by

$$\langle n_l \rangle = e\sum_{J=1}^{M} P_J \mathbf{n}_J \bullet \mathbf{u}_l$$

where $\mathbf{u}_l$ is a unit vector in the direction of node l.

To complete our description of quantum transport it is therefore necessary to determine the legal set of soliton states $\{\Psi\}$ and the probability density of those states P_J. Fortuitously, a further traffic model exists which describes the dynamic formation/destruction of the soliton states. Energy considerations show that the number of legal soliton states is *finite* for a given bias potential ($\{\Psi\} = \{\mathbf{n}_1,\mathbf{n}_2,...\mathbf{n}_M\}$; M=number of legal soliton states), and it is this finiteness which gives the real computational power to the traffic theoretic approach.

TRAFFIC THEORY II - SOLITON STATE LEVEL

Given a tunnel event between node k and node l there is a corresponding change in the electrostatic state $\Psi_I \rightarrow \Psi_J$ corresponding to $\mathbf{n}_I \rightarrow \mathbf{n}_J$, and we may define the state departure rate from $\Psi_I \rightarrow \Psi_J$ as $\mu_{IJ} = \Gamma(k,l;\mathbf{n}_I)$ (X_{kl} = 1) (see Fig. 5). The total departure rate is $\mu_I = \Sigma\, \mu_{IJ}$ (sum over J, $I \neq J$). For each visit, the system will dwell an average time $1/\mu_I$ in state Ψ_I. The transition probability from Ψ_I to Ψ_J is therefore $r_{IJ} = \mu_{IJ}/\mu_I$.We may collect the set of transition probabilities into a $M \times M$ routing matrix $\mathbf{R}$. The input traffic λ_I to state I is the superposition of the proportion of each departure-stream from all the other nodes so that

$$\lambda_I = \sum \lambda_J r_{IJ} \ , \quad \text{or } \boldsymbol{\lambda} \bullet (\mathbf{I} - \mathbf{R}) = 0 \ . \tag{6}$$

Equations (6) are the traffic equations (Kleinrock, 1976); their solution gives the average traffic entering or leaving each state. If the solution is unique one may simply find the soliton state occupancy probability as the average arrival rate multiplied by the average waiting time

$$P_J = \lambda_J/\mu_I \ . \tag{7}$$

However, we find that $\det|\mathbf{I} - \mathbf{R}| = 0$, so that there is no unique solution to the traffic equations (6) and we must resort to statistical methods. To construct an ensemble theory we define $\mathbf{m}(t) = (m_1,m_2,...,m_M)$ where m_I= the number of soliton systems found in state $\Psi = \mathbf{n}_I$ at time t. m_I is a binary random variable with the property $\Sigma m_I = 1$. The network of soliton states corresponds to a closed network of servers (Fig. 7) in classical traffic theory in which there is exactly one job trapped inside. Under these conditions Jacksons' theorem (for a recent account see Leon-Garcia, 1994) shows that the joint probability distribution of $\mathbf{m}(t)$ for a network in equilibrium is separable. Using an extension of this theorem we find $P(\mathbf{m}) = (1/G)\alpha_1(m_1)\alpha_2(m_2)...\alpha_M(m_M)$ where G is a normalisation constant (of the partition function) and where the $\alpha_I(m_I)$ depend on the traffic properties at soliton node I.

If we now let $\lambda^* = (\lambda_1^*, \lambda_2^*,...\lambda_M^*)$ be some non-zero solution of the traffic equations, then

$$\alpha_I(t=0)=1;\ \alpha_I(t) = \lambda_I^*/\mu I \ ,$$

$$G = \sum_{\{m\}} \prod_{I=1}^{M} \alpha_I(m_I) = \sum_J \frac{\lambda_J^*}{\mu I} \ . \tag{8}$$

It follows that the probability of finding the system in soliton state $\Psi_I = \mathbf{n}_I$ is

$$P_J = \frac{1}{G}\frac{\lambda_J^*}{\mu I} \ . \tag{9}$$

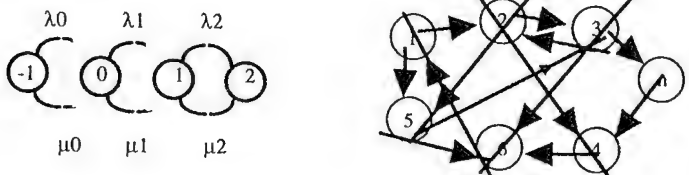


Figure 6. Soliton states in a 2-junction.

Figure 8. Transitions of soliton states

THE SINGLE JUNCTION

As a simple example of the queuing theory approach consider the 2-junction capacitor array (or single quantum dot) (Babiker and Barker, 1993). Above a threshold voltage V_{th} electrons can tunnel into and away from the dot, one electron at a time. If the dot has N excess electrons then the total electrostatic energy due to this charge is $(Ne)^2/2C_T$ where C_T is the total capacitance seen by the charge, $C_T = C_1+C_2+C_o$ and C_1, C_2 are the capacitances between the dot and the metallic electrodes and C_o is the capacitance to the ground electrode. In this case, on leaving state S_i the system can only make a transition to states S_{i+1} or S_{i-1} (see Fig. 6). This is a Markovian birth-death process which can be solved exactly using the transition rates λ_n ($S_{n-1} \rightarrow S_n$) and μ_n ($S_n \rightarrow S_{n-1}$) given in terms of the tunnelling rates by: $\lambda_n = \Gamma_{n-1}(l,c) + \Gamma_{n-1}(r,c)$; $\mu_n = \Gamma_n(c,l) + \Gamma_n(c,r)$. The probability of finding the system at state S_n is found as

$$P_n = \begin{cases} \frac{1}{Z}\prod_{i=k_1+1}^{n}\frac{\lambda_i}{\mu_i}, & n > k_1, \\ \frac{1}{Z}, & n = k_1, \end{cases} \tag{10}$$

where the partition function Z is given as

$$Z = 1 + \sum_{i=k_1}^{k_2}\prod_{n=k_1+1}^{i}\frac{\lambda_n}{\mu_n}, \tag{11}$$

and where k_1 and k_2 are the minimum and maximum possible number of excess electrons that can be accommodated on the dot. The states contributing to the process can be discovered according to the condition: if $\lambda_i > 0$ and $\mu_i > 0$ then both Ψ_i and Ψ_{i-1} are legal states. Figure 9 shows the differenial conductances G/G_0 and soliton densities $\langle n \rangle$ as a function of bias V and gate potential V_g which are compared with results obtained by the Monte Carlo method.

TIME EVOLUTION

Suppose the system is found in state $\Psi = n_K$ at $t = 0$ with initial densities $P_I(0) = \delta_{IK}$. At a later time t the system will be in a mixture of all possible legal states with densities $P_I(t)$. The probability of finding the system in state n_I at time $t+\Delta t$ is found as

$$P_I(t+\Delta t) = P_I(t)\left[1 - \sum \mu_{JI}(t)\Delta t\right] + \sum_J P_J(t)\mu_{JI}(t)\Delta t ,$$

which leads in the limit $\Delta t \rightarrow 0$ to the rate equation

$$\frac{\partial P_I}{\partial t} = \sum_J P_J \mu_{JI} - P_I \mu_I . \tag{12}$$

As an application, consider a single quantum dot described by the 2 soliton state $\{\Psi\} = \{n, n+1\}$, with birth and death coefficients λ and μ respectively. If the system is known to be in state Ψ_n at $t=0$, the rate equation (12) may be solved to give

$$P_n(t) = \frac{\mu + \lambda e^{-(\mu+\lambda)t}}{\lambda+\mu}, \quad P_{n+1}(t) = \frac{\lambda}{\lambda+\mu}\left[1 - e^{-(\mu+\lambda)t}\right], \tag{13}$$

which in the limit $t \rightarrow \infty$ recovers our earlier result $P_n(\infty) = \mu/(\lambda+\mu)$; $P_{n+1}(\infty) = \lambda/(\lambda+\mu)$.

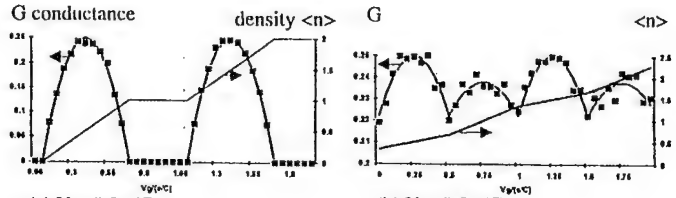


Figure 9. Conductances G/G_0 and soliton density $\langle n \rangle$ as a function of voltage V_g for an array of 2 tunnel junctions at two different bias voltages V. The dots are points calculated by Monte Carlo. The continuous lines denote the queuing theory results.

MORE COMPLEX SYSTEMS

The methodology easily encompasses multiple junction systems where it significantly outperforms the traditional Monte Carlo methods for determining the conductance (Babiker, 1994). The model has also been generalised to include *macroscopic quantum tunnelling* (Babiker and Barker, 1993) and the effects of *finite energy level spacing* (Babiker, 1994) suitable for 2DEG systems.

DISCUSSION AND CONCLUSIONS

The re-formulation of quantum transport theory of single electronic correlated-tunnelling systems in the form of traffic theory provides a powerful new representation for electron transport which is efficient when the number of different charge states is finite and the number of micro-events is countably finite. The formalism is readily extendable to a wide range of nanostructure problems but also has potential application in hopping theory and in the motion of solitonic excitations in polymer chains (Barker, 1994a). There are many improvements required if the theory is to be made more general. It is already obvious from the semi-classical "occupation number" formalism that we have described that a fully quantised theory (to include plasmons and the full range of excitations which contribute to image forces) should not alter the picture in too strong a way. An extension of the theory is needed clearly to bring out the details of the electrode relaxation and to overcome the assumption of the asymptotic limit which is implicit in our use of tunnelling rates. A first step in this direction has been made recently (Cluckie and Barker, 1994; Barker, 1994b) to determine the correlation dynamics of pairs of interacting electrons moving through polarisable tunnel junctions.The single biggest advantage however, in this new picture, is the ability to make analytical and non-perturbative analyses of the very complex transport problems inherent in single electronic systems. Some idea of the possible complexity in traffic networks is to be found in Braess paradox (Braess, 1968) which somewhat counter-intuitively essentially states that if a traffic network has relief roads added to it the net delays

may worsen! This feature of correlated transport may have important consequences for designing 1 bit on 1 electron logic circuits from single electronic systems.

REFERENCES

Averin, D. V., and Likhaerev, K. K., 1986, *J. Low Temp. Phys.* 62:345.
Babiker, S., 1994, Ph. D. Thesis, University of Glasgow.
Babiker, S., and Barker, J. R., 1993, in "Proc. Intern. Workshop on Computational Electronics," Ed. by C. Snowden, University of Leeds Press, 260.
Barker, J. R., 1994a, *in* "Introduction to Molecular Electronics," Ed. by M. Petty, D. Bloor, and M. Bryce, Edward-Arnold, London, 345.
Barker, J. R., 1994b, *Semiconduc. Sci. Technol.* 9:in press.
Barker, J. R., Roy, S., and Babiker, S., 1992a, *in* "Science and Technology of Mesoscopic Structures," Ed. by S. Namba, C. Hamaguchi, and T. Ando, Springer-Verlag, London, 213.
Barker, J. R., Weaver, J. M. R., Babiker, S., and Roy, S., 1992b, *in* "Proc. 2nd Intern. Symp. New Phenomena in Mesoscopic Structures," Ed. by C. Hamaguchi.
Braess, D., 1968, *Unternehmenforsch.* 12:258.
Cluckie, J., and Barker, J. R., 1994, *Semiconduc. Sci. Technol.* 9:in press.
Cooper, R. B., 1972, "Introduction to Queueing Theory," Macmillan Company, New York.
Fulton, T. A., and Dolan, G. J., 1987, *Phys. Rev. Lett.* 89:109.
Geerligs, L. J., Anderagg, V. F., Holweg, P. A. M., Mooij, J. e., Pothier, H., Esteves, D., Urbina, C., and Devoret, M. H., 1990, *Phys. Rev. Lett.* 64:2691.
Geerligs, L. J., Harmans, C. J. P. M., Kouwenhouven, L. P., Eds., 1993, "The Physics of Few Electron Nanostructures," *Physica B* 89:1.
Kelly, F. P., 1991, *Phil. Trans. Roy. Soc. (London) A* 337:343.
Kleinrock, L., 1976, "Queueing Systems," Academic Press, New York.
Kuzmin, L. S., Delsing, P., Claeson, T., and Likhaerev, K. K., 1989, *Phys. Rev. Lett.* 60:309; 62:2539.
Leon-Garcia, A., 1994, "Probability and Random Processes for Electrical Engineering," 2nd Ed., Addison-Wesley, New York, 546.
Likhaerev, K. K., 1988, *IBM J. Res. Develop.* 32:144.
Meirev, U., Kastner, M. A., and Wind, S. J., 1989, *Phys. Rev. B* 40:5871.
Weaver,???????

TRAJECTORIES IN QUANTUM TRANSPORT

John R. Barker

Nanoelectronics Research Centre
Department of Electronics and Electrical Engineering
University of Glasgow
Glasgow G12 8QQ, Scotland, UK.

INTRODUCTION

As device geometries approach nanometer scales the number of carriers in the active channel of a device or on the effective charging node of a single-electronic device approach very small numbers. Although such devices exist already they are generally measured by extracting a small but macroscopic current using techniques such as lock-in amplification. This is particularly true of scanning tunneling microscopy which may involve drawing current through a few-electron system over an effectively long period of time. The results are interpretable in terms of ensemble measurements on the system which is why in many cases it is possible to map out portions of the electron quantum probability distributions in ultra-small systems. However, the application potential of few-carrier devices and *a fortiori* siingle-electronic devices will depend on whether it will be possible to build a digital technology with one bit on one or a few carriers. If this information theoretic limit is reached it will be necessary to follow the discrete arrival of electrons in time and space at various points in the circuit; this will be similar to the quantum limit of optical communications - the shot noise regime.

There are two aspects to the discreteness of the electron which will become significant; the first is the need to count electrons, the second is to account for the stochastic nature of the electron position variable which is a direct consequence of quantum mechanics. This regime, where the graininess of charge (both itinerant and trapped) is significant has been called the granular limit (Barker, 1991, 1992a, 1992b, 1994; Barker *et al.*, 1992; Barker and Ferry, 1980); it is a regime where the once esoteric area of making measurements on single quantum systems becomes part of technology. From a device modelling point of view this regime already poses problems because one must abandon the comfortable self-consistent field calculations based on the simplified Posson equation where the inhomogeneous source term is given by the average carrier charge density. If the carriers were classical particles this would cause no major problems because one could in principle

track the particle trajectories via a Liouville equation or by direct Hamiltonian dynamics. Even classically, an analysis of device behaviour for few carriers is difficult because the granular limit is also a regime where strong carrier-carrier correlation is expected particularly through the direct Coulomb interaction. Quantum mechanically, the problem of describing individual electrons is still controversial even in the asymptotic limit. The question of how far we can describe electrons by trajectories is the subject of the present paper. A final answer is not given, instead it is hoped that the discussion will focus attention on a regime which has traditionally been the domain of the philosophy of quantum mechanics but which now could be exposed to direct experimental study.

ORTHODOX APPROACHES TO THE QUANTUM TRAJECTORY PROBLEM

Orthodox quantum mechanics has the semantic structure of a field theory (Bunge, 1967). As such the notion of a particle trajectory and consequently the concept of an arrival time appears not to be fundamental. Construction of transit times, tunnel times or times of arrival and their relation to trajectories and observable quantities must therefore be carried out at *secondary* level. Later we shall examine non-orthodox interpretations of quantum mechanics which permit *primary* levels of identification of trajectories. Three typical secondary approaches may be referred to as: (i) the dynamical approach; (ii) the asymptotic wavepacket model and (iii) the path variable quantum distribution approach.

(i) It has been remarked by many that time is not a dispersive variable in quantum mechanics and so a time of arrival operator and energy-time uncertainty relations are not possible. However, for those dynamical variables A which are time-dependent (even in the formal sense that $[H, A] \neq 0$), and for which an inverse A^{-1} exists we may define a "relaxation time" operator or t_Λ or equivalently a relaxation rate operator $t_\Lambda^{-1} = A^{-1}dA/dt$ (Barker and Ferry, 1980), which is a dispersive variable and for which one may derive "uncertainty relations" and so on. This dynamical construction has some merit in examining the issue of tunnel time. There is the intriguing possibility of finding a form for t_Λ which would commute with position operator thus formally defining a trajectory.

(ii) Although wavepackets may be constructed (and there is an important issue of the experimental preparation of a wavepacket) to be reasonably localised there is no inherent feature in a wavepacket which can unambiguously be associated with a feature for determining a time of arrival or a tunnel time or a precise trajectory. However for gaussian wavepackets which are incident on and scatter/tunnel from a localised tunnel barrier and for which the asymptotic forms are also gaussian it is possible to track trajectories associated with the packet peaks and/or centres of mass within particular regions. By extrapolating asymptotic trajectories backwards in time into the interaction zone one may build up a simple empirical picture of wavepacket tunnel time which may be used for comparison with analytical theories (Barker, 1985, 1986; Collins *et al.*, 1987, 1988). Of course this construction does not refer to the position of a particle, only to an asymptotically defined feature of a wavepacket. Extensive studies have been made of the concept of tunnel time for both reflected and transmitted waves (Büttiker and Landauer, 1986) many of the results of which are directly obtainable by Feynman path integral methods reviewed recently in (Jauho, 1992; Leavens and Aers, 1993). In most of these cases it is difficult to pick out a concept of trajectory.

(iii)The parameterisation of Wigner functions by path variables has often been suggested as a route to the derivation of quantum phase space trajectories. There are severe problems here, including those due to the non-compact suppport of Wigner distributions and the need to use a projected set of boundary conditions based on a complete set of Wigner functions which necessarily include complex terms (Barker, 1992b; Barker and Murray, 1983). An example of a set of trajectories is shown in Fig. 1; "contours" derive from the stationary Wigner distribution (Barker and Lowe, 1981).

Other parameterised trajectories are feasible (Bunge, 1967) for modelling quantum distributions mathematically or computationally, but as with the "trajectories" in Feynman path integrals they do not necessarily carry additional predictive power such as determining transit times. The wavepacket and quantum distribution trajectory approaches described above are deterministic (although bifurcation exist) since they are basically examining transit times from the viewpoint of the Schödinger equation which is deterministic. In contrast, the Born interpretation forces a stochastic viewpoint in which from a Copenhagenist viewpoint we might have some subjective probability of localising an electron in some initial region followed by a subjective probability of "finding" the electron in some other region at a later time. Some curious features occur in this scenario some of which have been tested experimentally and which point to the inadequacy of orthodox quantum mechanics as an interpretive framework, but more importantly expose regimes where quantum mechanics may be put at risk.

NON-ORTHODOX APPROACHES TO QUANTUM TRAJECTORIES

An unambiguous definition of a quantum trajectory is provided in the non-orthodox interpretation of quantum mechanics due principally to Bohm (1952) and de Broglie. In the Bohm approach a particle has a well-defined *deterministic* trajectory which is determined by the classical forces plus a quantum force which derives in a subtle non-local way from the wavefunction of the particle. If the wavefunction is known, a particle at location $\mathbf{r}$ is defined to have momentum $\mathbf{p} = \nabla S$, where S is the phase of the wavefunction.The full explanatory power of the Bohm theory was not evident until the computer simulations carried out between 1979 and 1987 (Philippidis et al., 1979a, 1979b; Dewdney and Hiley, 1982; Vigier *et al.*, 1987) where many classic 1D and 2D quantum mechanical time-dependent wavepacket problems such as two-slit interference, the Aharonov-Bohm effect and tunnelling through barriers were critically re-examined in the Bohm picture. Extensive reviews of this picture have been made recently (Holland, 1993; Bohm and Hiley, 1994). The concept of transit time or tunnel time is inherent in the Bohm picture and may be inferred directly from the data presented in references 1987 (Philippidis et al., 1979a, 1979b; Dewdney and Hiley, 1982; Vigier *et al.*, 1987; Bohm and Hiley, 1994). This possibility has been discussed in detail by a number of workers (Barker, 1992a, 1994; Barker *et al.*, 1992; Hirschfelder *et al.*, 1974; Leavens, 1990; Spiller *et al.*, 1990). Although the Bohm picture has considerable merit it also has serious interpretive problems which suggest that it is only partly able to give us reliable trajectory picture. Let us first review the useful features.

TRAJECTORIES AND TRANSIT TIMES IN THE BOHM PICTURE

The Bohm-de Broglie theory is a coupled field-particle theory in which a point particle with position $\mathbf{r}(t)$ is coupled to the Schrödinger *pilot-field* $\psi = R\exp(iS/\hbar)$ via the real polar components $R(\mathbf{x},t)$ and $S(\mathbf{x},t)$. $\rho = R^2$ gives the probability density for the position of the particle, whence its momentum is determined by the gradient of the phase at that point $\mathbf{p} = \nabla S$. The subsequent trajectory is deterministic and we may write it in terms of path variables s (Barker, 1994) for a stationary (open or closed) state in identical form to the ray equations of optics:

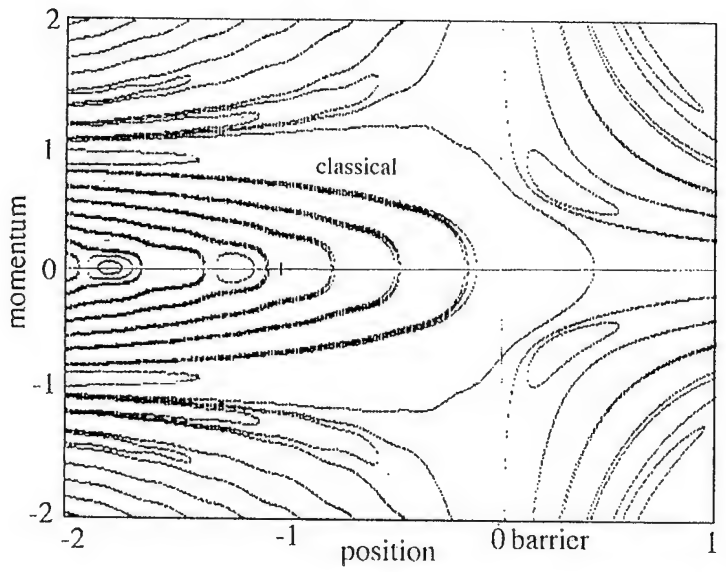


Figure 1. Trajectories in phase space derived from the stationary Wigner function for a plane wave scattering below the barrier height on a semi-infinite potential step in 1D.

$$\frac{d}{ds}\left(n\frac{d\mathbf{r}}{ds}\right)=\nabla n \quad , \tag{1}$$

$$n^2=\frac{\varepsilon - V - V_Q}{\varepsilon} \quad . \tag{2}$$

Here n is the analogue of a refractive index; it involves ε the total energy, V the potential energy and the quantum potential V_Q coupling to the pilot field:

$$V_Q=-\frac{\hbar^2}{2mR}\nabla^2 R \quad . \tag{3}$$

The controversial problem (Holland, 1993) of traversal time and in particular the concept of tunnelling time may be precisely established in the pilot field picture as first pointed out by Hirschfelder *et al.* (1974). The different concepts of dwell time, time of arrival, average traversal time may all be computed from the trajectories in the pilot-field picture. For example we may use the path variable $s(\mathbf{r},t)$ and the relationship $ds/dt = |\nabla S/m|$ to construct the traversal time between two points on a path:

$$t=\int ds\left|\frac{\nabla S}{m}\right|^{-1} \quad , \tag{4}$$

$$t=\int d\mathbf{r}\bullet\left(\frac{\nabla S}{m}\right)\left|\frac{\nabla S}{m}\right|^{-2} \quad . \tag{5}$$

The traversal time t may be regarded as a function of path length $t = t(s_1,s_2)$ where it is single valued, but expressed as a function of position $t = t(\mathbf{r}_1,\mathbf{r}_2)$ it is multiple-valued because a trajectory may re-visit the same location several times (for time dependent problems). Of course the exact initial location of a particle is rarely known, but in the Bohm picture as well as the orthodox picture the initial position is given with probability density r(x,t) which may be used to compute averaged versions of (4)-(5) . Thus the phase space distribution function for a Bohm particle is (Barker, 1992; Barker *et al.*, 1992)

$$f_B(\mathbf{r},\mathbf{p})=\rho(\mathbf{r},t)\delta\left(\mathbf{p}-\nabla S(\mathbf{r},t)\right) \quad . \tag{6}$$

As an example, the *average* dwell time within a region x_1 to x_2 for a 1D motion may be written (Leavens and Aers, 1993):

$$t_{Dwell}=\int dt\int_{x_1}^{x_2} dx\,\rho(x,t) \quad , \tag{7}$$

which coincides exactly with a standard Feynman path integral derivation (Sokolovski and Baskin, 1987). The result is obtained by defining the dwell time classically on the Bohm trajectory as

$$\begin{aligned} T_{Dwell} &= \int dt\,\theta\left(x(x_1,t)-x_1\right)\theta\left(x_2-x(x_1,t)\right) \\ &= \int dt\int dx\,\delta\left(x-x(x_1,t)\right)\theta\left(x-x_1\right)\theta\left(x_2-x\right) \end{aligned} \quad , \tag{8}$$

and then averaging over the initial distribution $\rho(x,t)$.

Unfortunately, this construction and in particular various statistical parameters such as the variance of momentum do not always give the correct quantum mechanical results without heuristically introducing other factors related to the quantum potential (Barker, 1992; Barker *et al.*, 1992). In particular the uncertainty relations fail .

PROBLEMS WITH THE BOHM PICTURE

The first problem with the pilot-field theory is the unsatisfactory use of a statistical boundary condition $\rho_{initial} = R^2(\mathbf{x},t)$ for a theory which is deterministic (both fields and particle evolve deterministically). It points up the immense degeneracy in the theory: the particle may be started anywhere in the field with probability density ρ. A second problem concerns an auxiliary requirement of the Bohm theory if it is to recover the Schrödinger equation, it is the the phase condition

$$S = S \bmod(\hbar) \quad . \tag{9}$$

It has no simple physical explanation but it leads to profound effects such as the non-integrable phase phenomenon in the Aharonov-Bohm effect. In the quantum hydrodynamic picture of quantum transport this condition necessarily leads to quantised vortices - the flow is no longer irrotational. This does not seem to have been appreciated in the device modelling literature, yet it includes the mechanism which gives rise to universal conductance fluctuations.

The de Broglie-Bohm trajectories have some peculiarities: (i) in a steady-state system the trajectories in configration space cannot cross in contra-distinction to purely classical theories where such an effect occurs only in phase space; (ii) in many stationary state problems the trajectories counter-intuitively reduce to points (ie zero momentum, for example in the particle in a box problem and most astonishingly in the case of a steady-state quantum stadium (2D) where classically one would have chaotic orbits); essentially the total energy resides in the quantum potential; (iii) relatively simple problems such as infinite plane waves incident on a tunnel barrier or potential in 1D are difficult to understand (by using extended wavepackets one finds however that some sense returns (Leavens, 1990) and the trajectories are reminiscent of ray trajectories in optics).

More seriously the relation between observable dynamical properties and the particle trajectories is no longer simple nor indeed local. For example, the mean kinetic energy of an ensemble of particles all in the same state should be given by

$$\left\langle \frac{p^2}{2m} \right\rangle = \int d^3\mathbf{r} R^2 \frac{(\nabla S)^2}{2m} \quad , \tag{10}$$

whereas it is actually given by

$$\left\langle \frac{p^2}{2m} \right\rangle = \int d^3\mathbf{r} R^2 \left[\frac{(\nabla S)^2}{2m} + V_Q \right] . \tag{11}$$

Provided the normal quantum construction

$$\langle A \rangle = \int d^3\mathbf{x} \Psi^*(\mathbf{x},t) A(\mathbf{x},\mathbf{x}') \Psi(\mathbf{x}',t) = \int d^3\mathbf{x} \rho(\mathbf{x},\mathbf{x}',t) A(\mathbf{x},\mathbf{x}') \tag{12}$$

is used for dynamical variables the de Broglie-Bohm theory automatically generates sensible averages, but the procedure is clearly inconsistent and the interpretation becomes very clouded. Supporters of the approach argue that any measurable quantity necessarily involves the local particle properties determined by r and p and the interaction with the field such as the appearance of V_Q in expression (11), but no consistent physical picture of how to do this other than via (12) has been forthcoming.

The de Broglie-Bohm picture leads automatically to the existence of a phase-space distribution function $f_B(\mathbf{r},\mathbf{p})$ given by (6). This positive-definite distribution gives the correct ensemble average for any classical variable which depends on position alone or is linear in momentum. Generally however, the appropriate quantities A_B which convolve with f_B to give the averages $\langle A \rangle$:

$$\langle A \rangle = \int d^3\mathbf{r} \int d^3\mathbf{p} A_B(\mathbf{r},\mathbf{p}) f_B(\mathbf{r},\mathbf{p}) \quad , \tag{13}$$

are not the corresponding classical functions of r and p; they are difficult to construct and worse are generally non-local operators as to be expected from a comparison of (12) and (13). A simple local example is

$$\langle p^2 \rangle = \int d^3\mathbf{r} \int d^3\mathbf{p} p_B^2(\mathbf{r},\mathbf{p}) f_B(\mathbf{r},\mathbf{p}) \;\; ; \;\; p_B^2(\mathbf{r},\mathbf{p}) = p^2 + 2mV_Q(\mathbf{r}) \quad . \tag{14}$$

There is an obvious conflict between conventional density matrix theory as represented by the *non-positive* definite Wigner phase space distributions $f_W(\mathbf{r},\mathbf{p})$ of many body quantum transport theory (or equivalent thermodynamic Green functions) and the *positive-definite* phase space probability distributions $f_B(\mathbf{r},\mathbf{p})$ implied by the existence of well-defined trajectories in the pilot field picture. Indeed, explicit expressions for the two in the case of a single electron pure state give:

$$f_W(\mathbf{r},\mathbf{p}) = \frac{1}{h^3} \int d^3\mathbf{x} R(\mathbf{r} - \frac{\mathbf{x}}{2}) R(\mathbf{r} + \frac{\mathbf{x}}{2}) e^{-i\mathbf{p}\cdot\mathbf{x}/\hbar} \exp\left[\frac{i}{\hbar} \left[S\left(\mathbf{r} + \frac{\mathbf{x}}{2}\right) - S\left(\mathbf{r} - \frac{\mathbf{x}}{2}\right) \right] \right] . \tag{15}$$

Unlike the Wigner function, the function $f_B(\mathbf{r},\mathbf{p})$ does have *compact support:* it exists precisely where the wavefunction exists. By contrast the Wigner function exists on the convex hull of the wavefunction. However, to obtain the correct statistical averages of dynamical variables great care is needed in taking account of the coupling of the electron to its pilot field (in fact a similar, but resolvable, problem occurs in the Wigner picture because a general classical variable $A(\mathbf{r},\mathbf{p})$ will lead to complicated counterpart $A_W(\mathbf{r},\mathbf{p})$ according to the Wigner-Weyl transformation). For example, in any measurement of the electron kinetic energy must be considered as measuring a particle embedded in the pilot field it is necessary to include the interaction energy V_Q with the pilot field as well as the kinetic term $p^2/2m$ in arriving at the observable kinetic energy. The Wigner function version of (14) gives

$$\langle p^2 \rangle = \int d^3\mathbf{r} \int d^3\mathbf{p} p_W^2(\mathbf{r},\mathbf{p}) f_W(\mathbf{r},\mathbf{p}) \;\; ; \;\; p_W^2(\mathbf{r},\mathbf{p}) = p^2 \quad . \tag{16}$$

These arguments may be used to derive the Heisenberg uncertainty relations or statistical dispersion relations arise in the two approaches (Barker, 1992, 1994; Barker *et al.*, 1992b). Identical results are obtained in the two approaches by remembering to compute the interaction energy with the pilot field; i.e., the quantum potential in determining the conventional kinetic energy. The interpretations are quite different however. In the orthodox approach, the uncertainty relations describe the statistical scatter in complementary observables. In the pilot wave picture the occupancy of a deterministic trajectory is determined randomly according to the initial position distribution. The variance of the momentum does not satisfy the Heisenberg relations. The Heisenberg relations in the pilot-field picture hold for the variance of an effective momentum given by (14). This is very unsatisfactory for deriving hydrodynamic device models from the Liouville equation for f_B and it is a very unsatisfactory conceptually as a particle picture. The validity of the de Broglie-Bohm approach and the existence of the coupling to V_Q could be tested by single-electronic analogues of the optical time of traversal experiments (Kwiat *et al.*, 1993; Steinberg *et al.*, 1994). In particular the de Broglie-Bohm theory predicts a precise arrival time which depends on final location and according to the deterministic trajectory picture there should be no statistical dispersion along a given trajectory.

Many of the above problems with the de Broglie-Bohm picture stem from the reluctance to give up the concept of a deterministic trajectory.

A STOCHASTIC PILOT FIELD PICTURE OF TRANSPORT

In a recent study (Barker, 1994), a deeper de-construction of the Schrödinger theory is put forward to gain a further pilot-field picture which has the merits of being self-consistent (the Bohm boundary condition is removed) and construction-reversible but in which the deterministic picture of particle trajectories is lost. Trajectories still occur but they are stochastic. A new definition of average tunnel time and a probability distribution of tunnel times is thereby proposed.

As a minimalist improvement on the de Broglie-Bohm theory we suppose that a particle trajectory exists but the particle motion $\mathbf{r}(t)$ is allowed to be stochastic such that the mean momentum is given by the gradient of the phase and the variance in momentum is given by an expression involving a diffusion current. In this picture the defining equations are :

$$\langle \delta(\mathbf{x}-\mathbf{r}(t)) \rangle = R^2(\mathbf{x},t) = \rho \quad , \quad \mathbf{p}_0 = \nabla S(\mathbf{x}) \ , \tag{17}$$

$$\langle p \rangle = \int d^3\mathbf{r}\rho\nabla S(\mathbf{x}) \quad , \quad \langle p^2 \rangle = \int d^3\mathbf{r}\{\rho[p_0^2 + p_1^2]\} \ , \tag{18}$$

$$-\rho\frac{\partial S}{\partial t} = \rho\frac{(p_0^2 + p_1^2)}{2m} + \rho V - \frac{\hbar}{2m}\nabla\bullet(\rho\mathbf{p}_1) \ , \tag{19}$$

$$\frac{\partial\rho}{\partial t} + \nabla\bullet\frac{\rho\mathbf{p}_0}{m} = 0 \ , \tag{20}$$

$$\nabla\bullet\frac{\rho\mathbf{p}_1}{m} = \frac{\hbar}{2m}\nabla^2\rho \ . \tag{21}$$

Here the quantity $\mathbf{p}_1$ is a measure of the fluctuations in the momentum and may be determined from the stationary Fokker-Planck equation (21) as $\rho\mathbf{p}_1/m = (\hbar/2m)\nabla\rho$ (a diffusion current; diffusion coefficient $\hbar/2m$). Again we must augment thse equations with $S = S \bmod(\hbar)$. The equations (17-21) are an alternative de-construction of Schrödinger's equation and we note that the quantum potential is replaced by the sum of a kinetic energy term and a residual potential which has the form of a divergence of a diffusion current:

$$\rho V_Q = \frac{p_1^2}{2m} - \frac{\hbar}{2m}\nabla\bullet(\rho\mathbf{p}_1) \ . \tag{22}$$

The second term in (22) vanishes when integrated over the confining volume.

This picture may be solved self-consistently for the usual expectation values. We cannot know the individual trajectories in this picture because the motion is diffusive. In this picture the Heisenberg relations hold exavtly as statistical scatter relations, the previously stationary particle states are replaced by statistically stationary states described by the steady-state Fokker-Planck equation (21), a good deal of the energy represented by the quantum potential appears as fluctuation energy (entirely so in the stationary case of a particle in a box). The initial value problem inconsistancy is removed. The major observable effect is on the predicted traversal times. For situations in which the fluctuations in momentum $|\mathbf{p}_1|$ are small (near the classical limit) we estimate a random time of arrival with a mean given by

$$\langle\tau\rangle = \int d\mathbf{r}\bullet\frac{\nabla S}{m}\left[\left(\frac{\nabla S}{m}\right)^2 + p_1^2\right]^{-1} \ . \tag{23}$$

This result approximates to the de Broglie-Bohm prediction but in general the distribution of times of arrival will depend on the underlying random processes which are hidden. A semantic analysis of the new equations shows that we are no longer dealing with a regular Hamilton-Jacobi like dynamics instead the Schrödinger theory de-constructs into an ensemble averaged hidden dynamics. The appearance of diffusion-like terms and Fokker-Planck like relations suggests a range of possible 'hidden varriable' models.

ON THE ORIGIN ON OF A STOCHASTIC PILOT FIELD

The foregoing arguments suggest that if a physically meaningful trajectory representation of quantum transport exists it must be fully stochastic. A possible model was outlined by us in (Barker, 1994) where the mysterious phase condition $S = S \bmod(\hbar)$ is used as a starting point. Suppose that the Bohm-de Broglie equations represent the statistical expectation values of appropriate dynamical variables in an underlying stochastic coupled particle-field theory (where in general we identify "particle" with a local (in space-time) dynamical degree of freedom to be treated as a classical dynamical quantity as in the Bohm-de Broglie theory but coupled to a stoachastic field). One possible clue to the nature of an underlying stochastic field is in the phase condition: $S(\mathbf{x},t) = S(\mathbf{x},t)\bmod(\hbar)$. Suppose that this condition represents an ensemble average of the local particle action $S'(\mathbf{r}, t)$ where $\mathbf{r}$ is a representative particle position vector. Denoting the ensemble average by <...>, we may generally obtain $S = S \bmod(\hbar)$ by the ansatz

$$S'(\mathbf{r},t) = S_0(\mathbf{r},t) + \sum\{\eta_i\theta(s(\mathbf{r},t) - s_i) + \mu_i\theta(t - t_i\} \ , \tag{24}$$

which defines the action as the sum of a deterministic part $S_0(\mathbf{x},t)$ and the sum of random impulses $\eta_i(\mathbf{r},t)$ and $\mu_i(t)$ of possible value $\pm\hbar$: $\eta_i(\mathbf{r},t) = \pm\hbar$; $\mu_i(t) = \pm\hbar$ delivered at random locations labelled i along the trajectory path variable $s(\mathbf{r},t)$ and at times t_i. The ensemble average of S' is then indeterminate up to a factor of $n\hbar$, where n is an integer. scale of the fluctuations in momentum. With suitable assumptions (Barker, 1994) on the stochastic fields $\eta_i(\mathbf{r},t)$ and $\mu_i(t)$ it is possible to recover (17-21) as ensemble averages of classical motion in stochastic fields with the action given by (24).

REFERENCES

Barker, J. R., 1985, Physica 134B:22.

Barker, J. R., 1986, *in* "Physics and Fabrication of Microstructures and Microdevices," Ed. by M. Kelly and C. Weisbuch, Springer-Verlag, 210.

Barker, J. R., 1991, *in* "Granular Nanoelectronics," Ed. by D. K. Ferry, J. R. Barker, and C. Jacoboni, Plenum, New York, 327.

Barker, J. R., 1992, *in* "Handbook on Semiconductors," Vol. 1, 2nd Revised Edition, Ed. by P. Landsberg, Elsevier-North Holland, Amsterdam, 1079.

Barker, J. R., 1994, *Semicond. Sci. Technol.* 9:in press.
Barker, J. R., and Ferry, D. K., 1980, *Sol.-State Electron.* 23:531.
Barker, J. R., and Lowe, D., 1981, unpublished.
Barker, J. R., and Murray, S., 1983, *Phys. Lett.* 93A:271.
Barker, J. R., Roy, S., and Babiker, S., 1992a, *in* "Science and Technology of Mesoscopic Structures," Ed. by S. Namba, C. Hamaguchi, and T. Ando, Springer-Verlag, London, 213.
Barker, J. R., Weaver, J. M., Babiker, S., and Roy, S., *in* "Proc. Intern. Symp. on New Phenomena in Mesoscopic Structures," Ed. by C. Hamaguchi.
Bohm, D., 1952, Phys. Rev. 85:166, 180.
Bohm, D., and Hiley, B. J., 1994, "The Undivided Universe: an Ontological Interpretation of Quantum Mechanics," Routledge, London.
Bunge, M., 1967, "Foundations of Physics," Springer-Verlag, London.
Büttiker, M., and Landauer, R., 1986, IBM J. Res. Develop. 30:451.
Carruthers, P., and Zachariesen, F., 1983, *Rev. Mod. Phys.* 55:245.
Collins, S., Lowe, D., and Barker, J. R., 1987, *J. Phys. C* 20:6213,
Collins, S., Lowe, D., and Barker, J. R., 1988, *J. Phys. C* 21:6233.
Dewdney, C., and Hiley, B. J., 1982, *Found. Phys.* 12:27.
Hirschfelder, J., Christopher, A. C., and Palke, W. E., 1974, *J. Chem. Phys.* 61:5435.
Holland, P., 1993, "The Quantum Theory of Motion," Cambridge University Press, Cambridge.
Jauho, A. P., 1992, *in* "Hot Carriers in Semiconductor Nanostructures-Physics and Applications," Ed. by J. Shah, Academic Press, New York, 121.
Kwiat, P. G., Steinberg, A. M., and Chiao, R. Y., 1993, *Phys. Rev. A* 47:R2472.
Leavens, C. R., 1990, *Sol. State Commun.* 74:923; 76:253.
Leavens, C. R., and Aers, G. C., 1993, *in* "Scanning Tunneling Microscopy III," Springer-Verlag Series in Surface Science 29:105.
Philippidis, C., Bohm, D., and Kaye, R. D., 1979a, *Il Nouvo Cim.* B71:75.
Philippidis, C., Dewdney, and Hiley, B. J., 1979b, *Il Nouvo Cim.* B52:15.
Sokolovski, D., and Baskin, L. M., 1987, *Phys. Rev. A* 36:4604.
Spiller, T. P., Clark, T. D., Prance, R. J., and Prance, H., 1990, *Europhys. Lett.* 12:1.
Steinberg, A. M., Kwiat, P. G., and Chiao, R. Y., 1994, *in* "Proc. XXVIIIth Rencontre de Moriand," Ed. by J. Tran Thanh Van, Editions Frontieres, Gif-sur-Yvette.
Vigier, J. P., Dewdney, C., Holland, P., and Kyprianids, A., 1987, *in* "Quantum Implications," Ed. by B. J. Hiley and F. D. Peat, Routledge, London.

LOCALIZED ACOUSTIC PHONONS IN LOW DIMENSIONAL STRUCTURES

N. A. Bannov[1], V. V. Mitin[1], and M. A. Stroscio[2]

[1] Department of Electrical and Computer Engineering,
Wayne State University Detroit, MI 48202
[2] U.S. Army Research Office, P.O.Box 12211,
Research Triangle Park, NC 27709-2211

INTRODUCTION

Modern microfabrication techniques allow the creation of new free-standing quantum nanostructures which attract considerable attention and are studied by several research groups. These structures are, in fact, solid plates (slabs) or rods (bars) connected to a solid substrate by a side of the smallest cross-section. The major feature of free-standing structures is that the smallest dimensions of the structures may be as small as a few interatomic distances. This attribute gives rise to new interesting physical phenomena and opens new possibilities for applications. First of all, the electrons (holes) in these structures are quantized. In fact, free-standing structures represent waveguides for electron waves which have features substantially different from more conventional quantum structures. Such waveguides may have very high potential energy barriers for electrons, so new effects related to hot but quantized electrons are possible. The phonon subsystem will also undergo significant modification. In this paper we will present a review of the experimental and theoretical results of other authors as well as results of our own investigation on the quantization of the acoustic phonons in free-standing structures.

Free-standing nanostructures have been fabricated in several laboratories and we will endeavor to give appropriate references. Free-standing quantum wells (FSQWs) made of various metals, such as Al, Ag, Au had been prepared by electron-beam evaporation or molecular-beam epitaxy on a cleaved NaCl substrate and by subsequently dissolving the substrate (Grimsditch *et al.*, 1987; Bahdra *et al.*, 1989). The thicknesses of the films were as thin as 200 A and typical areas of the surfaces were roughly 1 mm^2. Semiconductor GaAs and InGaAs FSQWs have been fabricated from spatially and compositionally modulated superlattices using standard lithographic techniques and selective etching (Williams *et al.*, 1992). In those structures FSQWs were suspended between two support posts and the quantum wells were parallel to the surface of the substrate, so they reminded bridges. Such structures had reproducible well widths from 80-200 A. The typical in-plane sizes of the FSQWs were 2.5×0.25 μm^2.

Results of successful fabrication of free-standing quantum wires (FSQWIs) and free-standing quantum dots (FSQDs) are reported in (Viswanath *et al.*, 1994; Yoh *et al.*, 1993; Hiruma *et al.*, 1993; Foad *et al.*, 1992; Tsutsui *et al.*, 1993) (see also references in the review Viswanath *et al.*, 1994). There are basically two different approaches to FSQWIs fabrication. In one of them the FSQWs are prepared in a manner similar to that used for FSQWs preparation with the difference that the widths of such FSQWs are small and constitute only several thicknesses; hence such quantum structures may be considered as FSQWIs (Yoh *et al.*, 1993). This method had been used to make InAs FSQWIs with widths of 2000 A to 7000 A and a thickness of 150 A. In another approach very long GaAs, InAs, and Si whiskers were grown on a GaAs substrate by metal-organic epitaxy (Viswanath *et al.*, 1994; Hiruma *et al.*, 1993; Canham, 1990). Quantum dots were obtained by reactive ion etching whiskers in a mixture of CH_4 and H_2 (Foad *et al.*, 1992; Tsutsui *et al.*, 1993). The whiskers grew in the <111> direction of the GaAs substrate, whatever the substrate orientation, and had typical lengths of 1-5 μm and diameters of 10-200 nm.

There are several possible applications of the free-standing structures. They may be used for probing the local properties of solids and there are several works where such possibilities were demonstrated (Grimsditch *et al.*, 1987; Bhadra *et al.*, 1989). Free-standing quantum structures may find applications as very sensitive sensors of forces or displacements in ways similar to those used for thin film sensors (Itoh and Suga, 1994). And there exists a variety of potential uses of free-standing structures for electronic and photonic applications, e.g. as low voltage field emitters, light emitting devices, mirrors for optical resonators (Viswanath *et al.*, 1994; Tsutsui *et al.*, 1993; Ho *et al.*, 1990; Ravi and Marcus, 1991). In Haruma *et al.*, (1993) the GaAs FSQWIs (whiskers) were grown with built-in *p-n* junctions and ohmic contacts were fabricated at both sides of the structures. The photoemission spectra as well as the photoluminescence spectra were studied for these FSQWIs. The photoemission spectra displayed red shifts which were not observed in bulk structures grown in the same conditions. A satisfactory explanation of this phenomenon has not been found. The photoluminescence spectra revealed a strong dependence on the orientation of the excitation light polarization with respect to the FSQWs axes. Si FSQWIs illuminated with green light emit red light and this effect may be used in Si based optoelectronics (Canham, 1990). Optical and transport properties of the FSQWIs were studied theoretically by Sanders and Chang (1992) and Sanders *et al.* (1993).

In this paper we will concentrate our main attention on acoustic phonons - related physical phenomena in free-standing nanostructures. The quantization of the acoustic phonon spectrum results in modifications of the acoustic phonon interactions with electrons and photons and manifests itself in electrical and optical measurements. In fact, the mentioned acoustic phonon quantization takes also place in quantum wells and quantum wires lying on the substrate or buried in substrate and we shall also focus on the consideration of the confined phonons. The spectra of the acoustic phonons in opaque metal FSQWs were investigated by Grimsditch *et al.* (1987) and Bhadra *et al.* (1989). The authors claimed that the Brillouin light scattering technique which they used would work to detect any acoustical mode which produces undulations at the surface. Quantized acoustic phonons were observed even in conventional AlAs-GaAs-AlAs quantum wells by the photothermal luminescence spectroscopy method (Chen *et al.*, 1993). The conductance of the AuPd quantum films and wires with widths of 200 A made on a silicon substrate was studied by Nabity and Wybourne (1992) and by Seyler and Wybourne (1992). The variation in the conductance as a function of the applied electric field has periodic peaks which authors have attributed to the electron interactions with confined acoustic phonons.

To describe quantitatively electron transport and optical properties of the quantum structures it is necessary to consider all of the acoustic phonon modes, their spectra, and their interactions with electrons and photons. Detailed understanding of confined acoustic

phonons in quantum structures and their spectra may also be significant for some of the nondestructive methods of diagnostic of microstructures where propagation of the acoustic phonons is employed (Challis *et al.*, 1990; Rosenfusser *et al.*, 1986; Lin *et al.*, 1993)

While there is an extensive literature on acoustic modes in acoustical waveguides, resonators and related structures (Mason, 1964; Auld, 1973), there are relatively few works considering this problem in a context of nanoscale structures (Wendler and Grigoryan, 1989, 1990; Sylla *et al.*, 1988; Velasko and Djafari-Rouhani, 1982; Akjouj *et al.*, 1987; Velasko and Garcia-Moliner, 1979; Kochelap and Gulseren, 1993; Kosevich and Khokhlov, 1968). In these papers, acoustic modes in systems with two interfaces are investigated and attention is drawn primarily to the modes localized between the interfaces. The peculiarities of acoustic phonon modes due to planar defects have also been considered (Kochelap and Gulseren, 1993; Kosevich and Khokhlov, 1968); it is shown that a few monolayers of different material (Kosevich and Khokhlov, 1968) or even a built-in electron sheet, interacting with phonons through the deformation potential (Kochelap and Gulseren, 1993), may result in localization of some acoustic modes on the planar defect. One-dimensional acoustical phonons in cylindrical free-standing quantum wires and their interactions with electrons were studied by several groups (Stroscio and Kim, 1993; Stroscio *et al.*, 1993; Grigoryan and Sedrakyan, 1983). The similar problems for FSQWs are considered in (Bannov *et al.*, 1993, 1994) and for FSQWIs of rectangular cross-section in (Kim et al., 1994; Morse, 1950).

In the following sections we will consider the acoustic modes in FSQWs. This may be determined analytically if we neglect the distortion of acoustic vibrations resulting from contact with the semiconductor substrate. This imposes restrictions on the in-plane wavelength, which should be shorter than a characteristic in-plane size of the semiconductor slab. These modes are normalized to introduce confined acoustic phonons. The acoustic phonon density of states is discussed.

EIGENMODES IN FREE-STANDING QUANTUM WELL

Small elastic vibrations of a solid slab can be described by a vector of relative displacement $\mathbf{u} = \mathbf{u}(\mathbf{r},t)$. Equations of motion of elastic continua have the form

$$\rho \frac{\partial^2 u_\alpha}{\partial t^2} = \frac{\partial \sigma_{\alpha,j}}{\partial r_j} \quad , \tag{1}$$

where ρ is the density of solid and $\sigma_{i,j}$ is the stress tensor. For isotropic continua

$$\sigma_{i,j} = \lambda u_{k,k}\delta_{i,j} + 2\mu u_{i,j} \quad ,$$

where λ, μ are the Lame constants, $u_{i,j}$ is the strain tensor,

$$u_{i,j} = 0.5\left(\partial u_i/\partial r_j + \partial u_j/\partial r_i\right) \quad ,$$

and $\delta_{i,j}$ is the Kronecker delta. The sum is assumed to be taken over repeated Roman subscripts. Equation (1) can be rewritten in a vector form as

$$\frac{\partial^2 \mathbf{u}}{\partial t^2} = s_t^2 \nabla^2 \mathbf{u} + (s_l^2 - s_t^2)\,\mathrm{grad}\,\mathrm{div}\,\mathbf{u} \quad , \tag{2}$$

where $s_l = (\lambda + 2\mu)/\rho$ and $s_t = \mu/\rho$ are the velocities of the longitudinal and transverse acoustic waves in bulk semiconductors.

We consider a solid slab of width a with free boundaries. Accordingly, the components of the stress tensor normal to the surfaces should be equal to zero. If we specify the coordinate system (which we will use throughout this paper) in such a way that the axis z is perpendicular to the semiconductor slab and the surfaces of the slab have coordinates $z=-a/2$ and $z=+a/2$, the boundary conditions take the form $\sigma_{x,z} = \sigma_{y,z} = \sigma_{z,z} = 0$ at $z=-a/2$ and $z=+a/2$; or in terms of components of the displacement vector, they have the following form

$$\begin{aligned} \sigma_{x,z} &= \mu\left(\partial u_x/\partial z + \partial u_z/\partial x\right) = 0 \quad , \\ \sigma_{y,z} &= \mu\left(\partial u_y/\partial z + \partial u_z/\partial y\right) = 0 \quad , \\ \sigma_{z,z} &= \lambda\,\mathrm{div}\,\mathbf{u} + 2\mu\left(\partial u_z/\partial z\right) = 0 \quad , \end{aligned} \tag{3}$$

at $z=-a/2$ and $z=+a/2$.

Our goal is to find eigenmodes for acoustic vibrations defined by (2) and (3). We will look for solutions in the following form

$$\mathbf{u}(\mathbf{r},t) = \sum_n \int \mathbf{u}_n(\mathbf{q}_t, z) \exp(i\mathbf{q}_t\mathbf{r}_t - i\omega_n t) \frac{d\mathbf{q}_t}{(2\pi)^2} \quad , \tag{4}$$

where $\mathbf{r}_t$ is the coordinate vector in the (x,y) plane, ω_n is the set of frequencies of vibrations. It can be proved that the eigenfunctions, $\mathbf{u}_n(\mathbf{q}_t, z)$, corresponding to nondegenerate eigenfrequencies, ω_n, are orthogonal. We can also orthogonalize eigenfunctions corresponding to equal eigenfrequencies using the Schmidt orthogonalization procedure. We will use $\mathbf{w}$ instead of $\mathbf{u}$ to denote the orthonormal set of eigenvectors, $\mathbf{w}_n(\mathbf{q}_t, z)$, defined by (2) - (4), for which

$$\int \mathbf{w}_n^+(\mathbf{q}_t, z)\mathbf{w}_m(\mathbf{q}_t, z)\,dz = \delta_{n,m} \quad . \tag{5}$$

The eigenvalue problem of (2)-(4) can be solved through the introduction of vector and scalar mechanical potentials which define the vector of relative displacement (Mason, 1964; Auld, 1973). The solution of this problem is known from acoustics and we will use acoustical terminology to identify eigenmodes. What has not been done in the field of acoustics is the normalization of eigenmodes (this is essentially a quantum mechanical problem); in addition, the peculiarities of phonon spectrum have not been investigated in detail. A major feature of the confined modes is their quantization in the z-direction. Roughly speaking, the z-components of the confined mode wave vectors, q_z, take only some discrete set of values at each particular in-plane wave vector, $\mathbf{q}_t$. There are three different types of confined acoustic modes: shear waves, dilatational waves and flexural waves. They are characterized by their distinctive symmetries. Here we will consider in detail the dilatational mode - the only one which interacts with electrons from the lowest electron subband (through the deformation potential). The formulae will be given in the Cartesian coordinate system with the axis x directed parallel to the vector $\mathbf{q}_t$, so $\mathbf{q}_t = (q_x, 0)$.

Dilatational waves

These waves are also called symmetric waves (in respect to the midplane) and have two nonzero components: $\mathbf{u}_n(q_l,z)=(u_x,0,u_z)$, where

$$u_x = iq_x\left[\left(q_x^2-q_t^2\right)\sin\left(\frac{q_t a}{2}\right)\cos(q_t z)+2q_l q_t \sin\left(\frac{q_l a}{2}\right)\cos(q_l z)\right], \tag{6}$$

$$u_z = q_l\left[-\left(q_x^2-q_t^2\right)\sin\left(\frac{q_t a}{2}\right)\sin(q_l z)+2q_x^2 \sin\left(\frac{q_l a}{2}\right)\sin(q_l z)\right]. \tag{7}$$

The parameters q_l, q_t are determined from the system of two algebraic equations

$$\frac{\tan(q_t a/2)}{\tan(q_l a/2)} = -\frac{4q_x^2 q_l q_t}{\left(q_x^2-q_t^2\right)^2}, \tag{8}$$

$$s_l^2\left(q_x^2+q_l^2\right) = s_t^2\left(q_x^2+q_t^2\right) \tag{9}$$

Equations (8) and (9) have many solutions for q_l and q_t at each particular q_x (Mason, 1964) and we label them by an additional index n : $q_{l,n}, q_{t,n}$. These solutions are either real or pure imaginary depending on q_x and n. We will use term *branches* of solutions to denote functions $q_{l,n}(q_x)$, $q_{t,n}(q_x)$, graphs of which are continuous single-connected curves. The frequencies of the dilatational waves are given by

$$\omega_n = s_l\sqrt{q_x^2+q_{l,n}^2} = s_t\sqrt{q_x^2+q_{t,n}^2}\ . \tag{10}$$

It is necessary to employ a numerical approach to solve the Eqs. (8) and (9). However, it is useful to make use of an analytical analysis initially in order to identify different branches and understand their general behavior.

If q_x=0, but $\omega_n \neq 0$, the roots of (8) and (9) may be obtain from the condition $\tan(q_t a/2)=0$. The appropriate solutions have the following form

$$q_t = \frac{2\pi n}{a}, \qquad q_l = \frac{s_t}{s_l}\frac{2\pi n}{a}, \qquad \omega = s_t q_t = s_l q_l, \qquad n = 1,2,3,\mathrm{K} \tag{11}$$

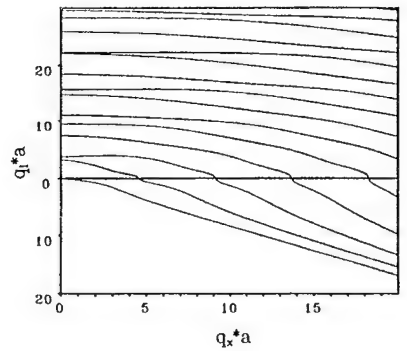
Each integer n from (11) identify a branch of solutions. The second set of branches may be obtained from the condition $\tan(q_l a/2)=\infty$ and in this case the solutions have the form

$$q_l = \frac{\pi+2\pi n}{a}, \qquad q_t = \frac{s_l}{s_t}\frac{\pi+2\pi n}{a}, \qquad \omega = s_t q_t = s_l q_l, \qquad n = 1,2,3,\mathrm{K} \tag{12}$$

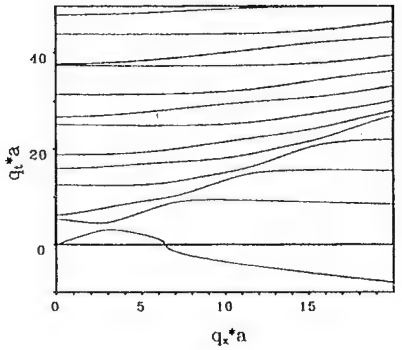
We will number branches by the integer index n in such manner that $\omega_n < \omega_{n+1}$ at q_x=0. We have to consider the case $q_x \to 0$ and $\omega \to 0$, which we have mentioned but not treated previously. From (10) it follows that both q_l and q_t should also go to zero. So, we may use the Taylor series expansion to obtain an approximate solution for the (8) and (9). The result is

$$q_l = i\frac{s_l^2-2s_t^2}{s_l^2}q_x, \qquad q_t = \frac{\sqrt{3s_l^2-4s_t^2}}{s_l}q_x, \qquad \omega = \frac{2s_t}{s_l}\sqrt{s_l^2-s_t^2}\,q_x.$$

We have a linear dispersion law for the lowest dilatational mode for small q_x and the velocity of this mode is smaller then s_l, but larger than s_t. An important peculiarity of this mode is the pure imaginary value of q_l, while the value of q_t is real. This means that the lowest dilatational mode contains terms $\sin(q_t z)$and $\cos(q_t z)$ which are extended throughout the width of the slab as well as terms $\sin(q_l z) = i\sinh(|q_l| z)$ and $\cos(q_l z) = \cosh(|q_l| z)$ which are localized at the surfaces of the slab [see (6) and (7)].



(a)



(b)

Figure 1. The solutions of the dispersion equation q_l (a) and q_t (b) as a function of q_x for dilatational phonons. The values of q_l and q_t above the abscissa are real and below the abscissa are pure imaginary.

The results of the numerical solution of Eqs. (6) and (7) are represented in Fig.1. The only material parameter which affects the solutions is the ratio of s_l / s_t. The numerical analysis was done for a GaAs slab with $s_l = 5.7\times10^5$ cm/s and $s_t = 3.35\times10^5$ cm/s. It is worth noting that values of q_l and q_t above the abscissa are real and below the abscissa are pure imaginary. It follows from the (9) that if for some q_x both q_l and q_t are real, then if q_x is increased, the value of q_t becomes pure imaginary while q_l remains real. The case where q_t is

pure imaginary and q_t is real is prohibited by (9) (in view of $s_t > s_l$). From the graphs on Fig.1 we may draw the conclusion that the $(q_t z)$ - dependent terms in the eigemodes turn into surface localized vibrations for large q_x whereas the $(q_t z)$ - dependent terms pass into surface localized vibrations for large q_x only for the lowest mode. The dispersion law for dilatational phonons, calculated for a 100 A width free-standing quantum well is given in Fig.2.

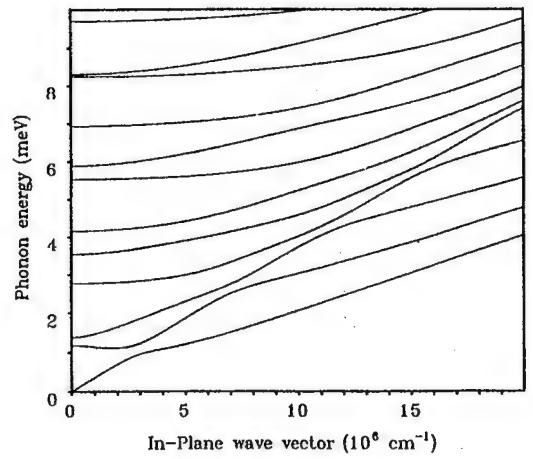


Figure 2. The dispersion law for dilatational phonons in a free-standing GaAs quantum well of width 10 nm for 12 lowest modes.

The flexural and shear modes may be analyzed in a manner similar to that presented for the dilatational modes. Their general behavior is roughly similar to the dilatational modes. However, unlike the dilatational modes, the flexural modes have antisymmetric displacement fields with respect to the midplane. They interact only with electrons from subbands of opposite parity (through the deformation potential). A very interesting feature of the lowest order flexural mode is the quadratic dispersion law which is similar to the widely known result for electrons. The shear modes do not interact with electrons through the deformation potential. They are similar to the bulk transverse modes and have only one displacement component which is perpendicular to both the direction of propagation and the normal vector to the quantum well.

Now we introduce the normalization constants $F_{s,n}$, $F_{d,n}$, and $F_{f,n}$, such that $\mathbf{w}_n = F_{s,n}\mathbf{u}_n$ for shear waves, $\mathbf{w}_n = F_{d,n}\,\mathbf{u}_n$ for dilatational waves, and $\mathbf{w}_n = F_{f,n}\,\mathbf{u}_n$ for flexural waves. Functions $\mathbf{u}_n$ are determined by (6) and (7) for dilatational waves and by similar expressions for shear and flexural waves (Bannov *et al.*, 1994).

The acoustic vibrations may be quantized in a standard manner using either the Lagrangian formalism (see e.g. Bannov *et al.*, 1994) or the principle that each normal mode carries energy $\eta\omega_n$ (see e.g. Stroscio and Kim, 1993). The operator for the relative displacement, $\mathbf{u}(\mathbf{r})$, may be represented through creation and annihilation operators, $c_n(\mathbf{q}_\parallel)$ and $c_n^+(\mathbf{q}_\parallel)$ (Bannov *et al.*, 1994) as

$$\mathbf{u}(\mathbf{r}) = \sum_{\mathbf{q}_\parallel, n} \sqrt{\frac{\eta}{2A\rho\omega_n(\mathbf{q}_\parallel)}} \left[c_n(\mathbf{q}_\parallel) + c_n^+(-\mathbf{q}_\parallel)\right] \mathbf{w}_n(\mathbf{q}_\parallel, z) \exp\left(i\mathbf{q}_\parallel \mathbf{r}_\parallel\right) . \qquad (13)$$

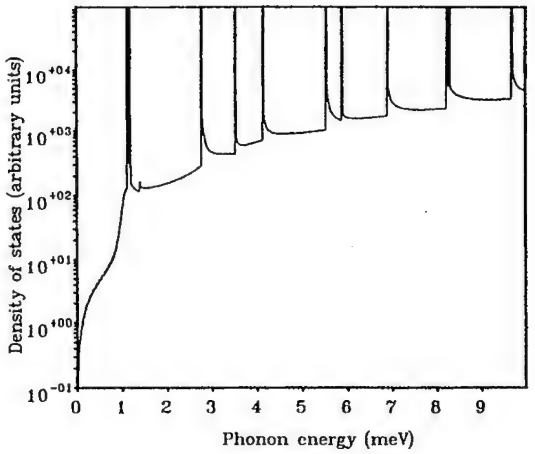


Figure 3. The density of states for dilatational phonons in a free-standing GaAs quantum well of width 10 nm.

ACOUSTIC PHONON DENSITY OF STATES

The peculiarities of the acoustic phonon spectrum will be markedly pronounced in the their density of states (DOS). The DOS of confined phonons is defined by the formula

$$N = \frac{A}{(2\pi)^2} \sum_n \int_{\omega_n = const} \left| d\omega_n / dq_\parallel \right|^{-1} dq_\parallel \ , \qquad (14)$$

where A is the area of the slab, and the sum is taken over phonon modes; integral in (14) is taken over the curve of constant energy and N is a function of the energy.

We have to specify the Brillouin zone to calculate the DOS over a wide range of energy. For a model estimation we accepted a simple square Brillouin zone. So we take into account only those acoustic phonons in integral (14) which have wavevectors inside the first Brillouin zone. The lattice constant is taken equal 5.65 A which corresponds to the case of GaAs. The graph of the DOS obtained by numerical calculation of the integral of (14) for dilatational phonons is depicted in the Fig. 3. At energies lower than some critical energy (corresponding to the edge of the Brillouin zone) the DOS is, on the average, a quadratic function of energy. This functional dependence occurs when many phonon branches contribute to the DOS and it corresponds to the case of bulk acoustic phonons. It is

obscured in Fig.3 because the graph is plotted in the semilogarithmic scale to emphasize the singularities of the DOS. These singularities correspond to the extrema in the dispersion relation; formally the DOS goes to infinity in such points. In Fig.3, the DOS is plotted for energies up to 10 meV. At higher energies the finiteness of the Brillouin zone becomes important and the function N saturates in the average.

The DOS may be determined experimentally from neutron scattering spectra (Bruesch, 1986; Di Bartolo and Powell, 1976) or from Brillouin light scattering spectra (Grimsditch *et al.*, 1987; Bhadra *et al.*, 1989). It is a very important function characterizing the acoustic phonon subsystem and determining peculiarities of phonon interactions with phonons, photons and electrons.

CONCLUSIONS

As a result of confinement acoustic phonons in free-standing structures undergo significant modifications which strongly affect the electrical and optical properties of such structures. Though the main features of the quantized phonons in low dimensional structures are clear, the details of their spectra and displacement fields have defied analytical analysis in structures which lack a high order of symmetry. Even a rectangular rod may be analyzed only if some model assumptions are made. In our view numerical analyses of the displacement fields would be useful for some typical structures; such results would facilitate future quantitative analyses of the scattering rates of electrons and photons by confined acoustic phonons in a wide variety of structures.

ACKNOWLEDGMENTS

We are grateful to V. Aristov for the performance of the computer calculations and preparation the graphs for this paper. This work was supported by ARO and NSF.

REFERENCES

Akjouj, A., Sulla, B., Zielinski, P., and Dobrzynski, L., 1987, *J. Phys. C*, 20:6137.
Auld, A., 1973, "Acoustic Fields and Waves," Wiley, New York (1973).
Bannov, N., Mitin, V., and Stroscio, M. A., 1993, *in* "Proc. Int. Sem. Dev. Res. Symp.," Charlottesville, 659.
Bannov, N., Mitin, V., and Stroscio, M. A., 1994, *Phys. Stat. Sol. (b)*, 183: in press.
Bhadra, B., Grimsditch, M., Schuller, I., and Nizzoli, F., 1989, *Phys. Rev. B*, 39:12456.
Bruesch, P., 1986, "Phonons: Theory and experiments II", Springer-Verlag, NY.
Canham, L. T., 1990, *Appl. Phys. Lett.*, 57:1046.
Challis, L. J., Kent, A. J., and Rampton, V. W., 1990, *Semicond. Sci. Technol.*, 5:1179.
Chen, Y. F., Chen, J. L., Lin, L. Y., and Huang, Y. S., 1993, *J. Appl. Phys.*, 73:4555.
Di Bartolo, B., and Powell, R., 1976, "Phonons and resonances in solids," John Wiley and Sons, New York.
Foad, M. A., Wilkinson, C. D., Dunscomb, C., and Williams, R. H., 1992, *Appl. Phys. Lett.*, 60:2531.
Grigoryan, V. G., and Sedrakyan, D. E., 1983, *Sov. Phys. Acoust.*, 29:281.
Grimsditch, M., Bhadra, R., and Schuller, I., 1987, *Phys. Rev. Lett.*, 58:1216.
Hiruma, K., Yazawa, M., Haraguchi, K., Ogawa, K., Katsuyama, T., Koguchi, M., and Kakibayashi, H., 1993, *J. Appl. Phys.*, 74:3162.
Ho, S. T., McCall, S. L., Slusher, R. E., Pfeiffer, L. N., West, K. W., Levi, A. Blonder, G., and Jewell, J., 1990, *Appl. Phys. Lett.*, 57:1387.
Itoh, T., and Suga, T., 1994, *Jpn. J. Appl. Phys., Pt.1*, 33:334.
Kim, K. W., Yu, S. G., Erdogan, M. U., Stroscio, M. A., and Iafrate, G. J., *in* "Proc.of the SPIE OE/LASE: Ultrafast Phenomena in Semiconductors."
Kochelap, V. A., and Gulseren, O., 1993, *J.Phys. Condens. Matter*, 5:589.
Kosevich, A., and Khokhlov, V., 1968, *Sov.Phys. - Solid State*, 10:39.
Lin, H.-N., Maris, H. J., and Freund, L. B., 1993, *J. Appl. Phys.*, 73:37.
Mason, W., Ed., 1964, "Physical Acoustics", Vol.1, Part A, Academic Press, New York.
Morse, R. W., 1950, *J. Acoustic. Soc. America*, 22:219.
Nabity, J. C., and Wybourne, M. N., 1991, *Phys. Rev. B*, 44:8990.
Ravi, T. S., and Marcus, R. B., 1991, *J. Vac. Sci. Technol. B* 9:2733.
Rosenfusser, M., Koster, L., and Dietsche, W., 1986, *Phys. Rev. B*, 34:5518.
Sanders, G. D., and Chang, Y. C., 1992, *Appl. Phys. Lett.*, 60:2525.
Sanders, G. D., Stanton, C. J., and Chang, Y. C., 1993, *Phys. Rev. B*, 48:11067.
Seyler, J., and Wybourne, M. N., 1992, *Phys. Rev. Lett.*, 69:1427.
Stroscio, M. A., and Kim, K. W., 1993, *Phys. Rev. B.*, 48:1936.
Stroscio, M. A., Iafrate, G. J., Kim, K. W., Mitin, V., and Bannov, N., 1993, *in* "Proc. Int. Sem. Dev. Res. Symp.," Charlottesville, 873.
Sylla, B., More, M., and Dobrzynski, L., 1988, *Surf. Sci.*, 206:203.
Tsutsui, K., Hu, E. L., and Wilkinson, C., 1993, *Jpn. J. Appl. Phys., Pt.1*, 32:6233.
Velasko, V., and Garcia-Moliner, F., 1979, *Physica Scripta*, 20:111.
Velasko, V., and Djafari-Rouhani, B., 1982, *Phys. Rev. B*, 26:1929.
Viswanath, A. K., Hiruma, K., Yazawa, M., Ogawa, K., and Katsuyama, T., 1994, *Microw. and Opt. Techn. Lett.*, 7:94.
Wendler, L., and Grigoryan, V. G., 1989, *Surf. Sci.*, 213:588.
Wendler, L., and Grigoryan, V. G., 1990, *Phys. Rev. B*, 42:1833.
Williams, M. D., Shunk, S. C., Young, M. G., Doctoer, D. P., Tennant, D. M., and Miller, B. I., 1992, *Appl. Phys. Lett.*, 61:1353.
Yoh, K., Nishida, A., Kunitomo, H., Ogura, T., and Inoue, M., 1993, *Jap. J. Appl. Phys. Pt.1*, 32:6237.

QUANTUM KINETICS IN LASER PULSE EXCITED SEMICONDUCTORS

H. Haug, K. El Sayed and L. Bányai

Institut für Theoretische Physik
J.W. Goethe Universität Frankfurt
Robert-Mayer Str. 8
D-60054 Frankfurt, Germany

INTRODUCTION

Laser light pulses can be as short as 6 fs = 6×10^{-15} s. Similarly fast are optical detection systems. The kinetics of the optically excited carriers can be monitored on a femtosecond time scale. In this regime the semiclassical Boltzmann kinetics breaks down and has to be replaced by a quantum kinetics. Thus quantum kinetics is needed for optics on a femtosecond time scale in the same way as for transport in nanometer microstructures, where the transit times are comparably short. In both cases the excited carriers behave, at least to some extent, like coherent quantum-mechanical waves. In low-dimensional microstructures with quantum confinement the phase space for phase-destroying scattering processes is considerably reduced, so that the coherent behavior of the electronic excitations is more long-lived than in bulk materials. For the description of the partially coherent behavior of the optically excited carriers, quantum kinetics is required.

We will treat the optical transitions across the fundamental energy gap of a semiconductor. For this purpose we have to include at least two bands, a valence band and a conduction band. Transitions between these bands are called interband transitions. Consider as the simplest case a semiconductor in which the extrema of both bands occur at the center of the Brillouin zone, i. e. at the momentum $k = 0$ and assume that the optical transitions with the absorption or emission of one photon are symmetry allowed. The absorption of one photon excites one electron from the valence band to the conduction band. Instead of treating $10^{23}-1$ electrons in the valence band, it is often wise to concentrate on the one missing electron in the valence band, called the hole. Thus the elementary optical excitation in a semiconductor creates an pair of oppositely charged quasi-particles, namely an electron in the conduction band and a hole in the valence band. The description of this optically induced interband kinetics requires, as in a two-level atom, the population densities of the upper and lower level, but also the polarization due to the coherent superposition the two states which is induced by the action of a coherent light field. In other words we have to calculate the reduced 2×2 density matrix for each k-state

$$\begin{bmatrix} \rho_{cc,k}(t) & \rho_{cv,k}(t) \\ \rho_{vc,k}(t) & \rho_{vv,k}(t) \end{bmatrix}, \tag{1}$$

where $\rho_{ij,k} = < a^{+}_{j,k} a_{i,k} >$. Here, $a_{i,k}$ is the annihilation operator of an electron in band i with momentum $\mathbf{k}$.

One can indeed treat the interband quantum kinetics in terms of reduced density matrices (Zimmermann, 1990, 1992; Mukamel, 1990; Schlösser *et al.*, 1992; Axt and Stahl, 1993; Schilp *et al.*, 1993) or in terms of non-equilibrium Green functions (Kusnetzov, 1991; Hartmann and Schäfer, 1992; Haug, 1992; Bányai *et al.*, 1992; Haug and Ell, 1992; Tran Thoai and Haug, 1992, 1993; El Sayed *et al.*, 1994a, 1994b; Kalvová and Velický, 1994). Both methods have their own advantages and also shortcomings. With the equation-of-motion technique one generates for the reduced density matrices a coupled set of differential equations for expectation values of an increasing number of operators, starting with the single-particle density matrix (expectation value of two electron operators at equal times). Naturally, one has to terminate this system of equation by a factorization of higher-order expectation values. The coupled differential equations which contain quantum correlation and non-Markovian memory effects implicitly, are well-suited for numerical solutions, but these effects cannot easily be displayed explicitly, nor does this technique allow simply to include certain effects up to infinite order in the perturbation, as, e.g., energy renormalizations, screening of interaction potentials, etc.. The non-equilibrium Green function theory, on the other hand, works only with single-particle propagators, but uses non-perturbational evaluations of the self-energies, which contain partial summations to infinite order. This technique results in integro-differential equations, which display quantum correlation and memory effects explicitly. However these integro-differential equations are more difficult to be solved numerically. In fact, the solution of these equations simplifies considerably if the integral kernels can be factorized. In this case one can also represent the integro-differential equations as a larger set of differential equations. We will formulate in terms of non-equilibrium Green functions as an introduction the interband kinetics for free-carriers, followed by a treatment of the interband quantum kinetics with LO-phonon interaction and finally with Coulomb interaction. The theory of Coulomb interaction is complicated by the fact that the screening of the Coulomb potential obeys its own nontrivial quantum kinetics. We will present a theory of the build-up of the screening and develop a particularly useful time-dependent plasmon pole approximation for the screened Coulomb potential. We show that the quantum kinetics of the early phase after a short femtosecond pulse is governed by a scattering kinetics without screening and without the validity of energy conservation for the individual collisions. The resulting optical dephasing agrees with a semiclassical treatment of the phase decay by Gurevich *et al.* (1991). The interband quantum kinetics with impurity scattering has been analysed in a recent paper by Kalvová and Velický (1994). This problem deserves attention because it is simple enough so that many analytical results can be obtained at least in the coherent potential approximation. A more detailed discussion of the quantum kinetics of laser pulse excited semiconductor is contained in a forthcoming book "Quantum Kinetics for Transport and Optics in Semiconductors" by Haug and Jauho (1994).

In the following we will limit ourselves, for simplicity, to a classical description of the laser-light field, but the photon kinetics can, in principle, also be included in a full quantum kinetic theory of the photons and the electrons in a semiconductor. The basic interaction of the electric field $E(t)$ of laser light with the valence electrons of a semiconductor is a dipole interaction. The spatial variation of the light field can often be neglected because the wavelength is usually much larger than other relevant length scales, e.g. the crystal cell. For the laser light field we will assume the form

$$E(t) = E_0(t)\cos(\omega t) \ , \tag{2}$$

where ω is the central frequency of the pulse, the time-dependent amplitude $E_0(t)$ will be taken usually in the form of a Gaussian pulse with a width τ and a peak amplitude E_0

$$E_0(t) = E_0 \exp\left[-\left(\frac{t-t_0}{\tau}\right)^2\right] . \tag{3}$$

Often we take into account only the resonant terms of the interaction, namely

$$H_I \sim -\frac{1}{2}\sum_k E_0(t)\left(d_k a^+_{c,k} a_{v,k} e^{-i\omega t} + \text{h.c.}\right) . \tag{4}$$

This is the so-called *rotating-wave approximation*. d_k is the optical dipole moment between the Bloch states in the valence and conduction band. In the vicinity of the Γ point one can disregard the k-dependence. If not absolutely necessary, we suppress the vector notation in the following. In the vicinity of the Γ point the energies of the electrons can be described by the effective mass m_e of the electrons in the conduction band and m_h of the holes in the valence band ($\hbar$=1 is used throughout for convenience):

$$e_{c,k} = \frac{E_g}{2} + \frac{k^2}{2m_e} = \frac{E_g}{2} + \varepsilon_{e,k} \ , \tag{5}$$

and

$$e_{v,k} = -\frac{E_g}{2} - \frac{k^2}{2m_h} = -\frac{E_g}{2} - \varepsilon_{h,k} \ . \tag{6}$$

For more details we refer to the text book on "Quantum Theory of the Optical and Electronic Properties of Semiconductors" by Haug and Koch (1993).

OPTICAL FREE-CARRIER INTERBAND KINETICS

As an introductory example we will disregard all interactions of the excited electrons and holes, and treat the free-carrier interband kinetics of a two-band direct-gap semiconductor. This simple model will be shown to lead to the optical Bloch equations of a semiconductor which will serve as an important paradigm in the whole field of the ultrafast kinetics of laser-pulse excited semiconductors. The relevant non-equilibrium electron Green functions in a spatially homogeneous crystal are now two-by-two matrices in the band index

$$G_{\mu\nu,k}(t_1,t_2) = -i\langle T_c[a_{\mu,k}(t_1)a^+_{\nu,k}(t_2)]\rangle \ , \tag{7}$$

with $(\mu,\nu) = (c,v)$. The time arguments and the time-ordering operator T_c are defined on the contour C. The self-energy of this non-equilibrium electron Green function can be written as

$$\Sigma_{\mu\nu,k}(t_1,t_2) = \Sigma^{\delta}_{\mu\nu,k}(t_1,t_2) + \theta(t_1-t_2)\Sigma^{>}_{\mu\nu,k}(t_1,t_2) + \theta(t_2-t_1)\Sigma^{<}_{\mu\nu,k}(t_1,t_2) \ . \tag{8}$$

The singular part of the self-energy Σ^{δ} has to be treated separately, as discussed by Danielewicz (1984). It is given by

$$\Sigma^{\delta}_{\mu\nu,k}(t_1,t_2) = \Sigma^{r}_{\mu\nu,k}(t_1)\delta(t_1-t_2) \ . \tag{9}$$

With the self-energy (9), the equation of motion for the particle Green function matrix becomes:

$$G_0^{-1}G^{<} = \Sigma^{\delta}G^{<} + \Sigma^{r}G^{<} + \Sigma^{<}G^{a} \ . \tag{10}$$

The corresponding r.h.s. differential form of the Dyson equation is

$$G^{<}G_0^{-1} = G^{<}\Sigma^{\delta} + G^{r}\Sigma^{<} + G^{<}\Sigma^{a} \ . \tag{11}$$

Subtracting the two equations of motion and putting $t_1 = t_2 = t$ yields

$$[G_0^{-1},G^{<}]_t = [\Sigma^{\delta},G^{<}] + [\Sigma^{r},G^{<}] + [\Sigma^{<},G^{a}] \ . \tag{12}$$

We start by treating as an introductory example free carriers only. Naturally many-body techniques are not required for this simple case, elementary equation of motion techniques are sufficient. In order to get some practice with the many-body techniques in the multi-band situation, we treat also this case with the non-equilibrium Green function method. The only remaining interaction in this problem is that of the electrons with the light pulse. It gives raise to a singular self-energy in the form

$$\Sigma^{r}_{\mu\nu,k}(t) = -dE(t)(1-\delta_{\mu\nu}) \ , \tag{13}$$

while the non-singular parts of Σ are zero. With (13) the equation of motion for $G^{<}_{\mu\nu,k}(t,t)$ becomes explicitly

$$\left[i\frac{\partial}{\partial t} - \left(e_{\mu,k} - e_{v,k}\right)\right]G^{<}_{\mu\nu,k}(t,t) = -dE(t)\sum_{\rho}\left[\left(1-\delta_{\mu\rho}\right)G^{<}_{\rho\nu,k}(t,t) - \left(1-\delta_{\rho\nu}\right)G^{<}_{\rho\nu,k}(t,t)\right] . \tag{14}$$

First, we see that in this simple case a closed equation for the reduced density matrices $G^{<}_{\mu\nu,k}(t,t)$ results. The diagonal elements of the density matrix simply define the electron densities in state i,k

$$G^{<}_{\mu\mu,k}(t,t) = i\left\langle a^+_{\mu,k}(t)a_{\mu,k}(t)\right\rangle = i\rho_{\mu\mu,k}(t) = in_{\mu,k}(t) \ . \tag{15}$$

$\rho_{\mu\mu,k}$ is the diagonal element of the density matrix

$$\rho_{\mu\nu,k}(t) = \left\langle a^+_{\mu,k}(t)a_{\nu,k}(t)\right\rangle \ . \tag{16}$$

Electron densities can be measured optically, e.g., with linear gain and absorption spectroscopy or with luminescence spectroscopy. However, in order to obtain the electron

distributions from such spectra, one always needs some theoretical analysis in terms of a line shape theory. The off-diagonal elements $\mu \neq \nu$

$$G^{<}_{\mu\nu,k}(t,t) = i\left\langle a^{+}_{\nu,k}(t) a_{\mu,k}(t)\right\rangle = i\rho_{\mu\nu,k}(t) \tag{17}$$

determine the interband polarization. $\rho_{cv,k}$, e.g. , describes the annihilation of a conduction band electron and the creation of a valence band electron. The relation

$$\rho_{\mu\nu,k}(t) = \left\langle a^{+}_{\mu,k}(t) a_{\nu,k}(t)\right\rangle = \rho^{*}_{\mu\nu,k}(t) \tag{18}$$

holds. The physical polarization of the medium by the light beam is obtained as

$$P(t) = \sum_{k} d_k \left[\rho_{vc,k}(t) + \rho_{cv,k}(t)\right] . \tag{19}$$

The measurement of the coherent polarization requires techniques of nonlinear optical spectroscopy, e.g. time-resolved four-wave mixing. In such an experiment the incoming beams create a laser-induced lattice. From the light diffracted from this lattice the polarization can be deduced.

In terms of the polarization and the densities we get the optical interband kinetic equations:

$$\left[\frac{\partial}{\partial t} + i\left(e_{c,k} - e_{v,k}\right)\right]\rho_{cv,k}(t) = -idE(t)\left[\rho_{cc,k}(t) - \rho_{vv,k}(t)\right] , \tag{20a}$$

$$\frac{\partial}{\partial t}\rho_{cc,k}(t) = -\frac{\partial}{\partial t}\rho_{vv,k}(t) = -idE(t)\left[\rho_{cv,k}(t) - \rho_{vc,k}(t)\right] . \tag{20b}$$

The complex conjugate of (20a) yields the equation for $\rho_{vc,k}$. The light field is of the form

$$E = E_0(t)\cos(\omega t) . \tag{21}$$

Introducing the electron-hole pair energy

$$e_{c,k} - e_{v,k} = e_k , \tag{22}$$

with

$$e_k = E_g + \varepsilon_k , \quad \text{with } \varepsilon_k = \varepsilon_{e,k} + \varepsilon_{h,k} = \frac{k^2}{2\mu} , \quad \frac{1}{\mu} = \left(\frac{1}{m_e} + \frac{1}{m_h}\right), \tag{23}$$

and the last quantity is the reciprocal of the reduceded electron-hole mass and only the resonant terms are taken into account, so that (rotating-wave approximation) we get, e.g.,

$$\begin{aligned} e^{i\omega t}\left(\frac{\partial}{\partial t} + ie_k\right)\rho_{cv,k}(t) &= \left[\frac{\partial}{\partial t} - i\left(\omega - e_k\right)\right]\rho_{cv,k}(t)e^{i\omega t} \\ &= -\frac{\rho_{cv,k}(t)}{T_2}e^{i\omega t} - \frac{\omega_R(t)}{2}\left[\rho_{cc,k}(t) - \rho_{vv,k}(t)\right] , \end{aligned} \tag{24a}$$

and

$$\frac{\partial}{\partial t}\rho_{cc,k}(t) = -\frac{\rho_{cc,k}(t)}{T_1} - i\frac{\omega_R(t)}{2}\left[\rho_{cv,k}(t)e^{i\omega t} - \rho_{vc,k}(t)e^{-i\omega t}\right] , \tag{24b}$$

where $\omega_R(t) = dE_0(t)$ is the Rabi frequency. In order to get a first phenomenological understanding of the influence of relaxation processes we included the relaxation times T_1 for the densities and T_2 for the polarization simply by hand. T_1 and T_2 are called the longitudinal and transversal relaxation time, respectively. The relaxation term for the electron density of the valence band is $-(\rho_{vv,k} - 1)/T_1$, because this density relaxes towards the full valence band. Naturally it will be the mature task of quantum kinetics to replace these simple relaxation terms by a more appropriate description of the relaxation processes.

The Optical Free-Carrier Bloch Equations.

Now one can introduce the three real components of the Bloch vector $\mathbf{U}_k$

$$U_{1k} = \rho_{cv,k}(t)e^{i\omega t} + \text{c.c.} , \tag{25a}$$

$$U_{1k} = i[\rho_{cv,k}(t)e^{i\omega t} - \text{c.c.}] , \tag{25b}$$

The first two components refer to the real and imaginary part of the microscopic polarization after elimination of the oscillations with the central frequency of the light pulse. The third component is the inversion. With the detuning

$$\delta_k = e_k - \omega , \tag{26}$$

one gets the following equations of motion

$$\frac{\partial U_{1k}}{\partial t} = -\Gamma_{11}U_{1k} - \delta_k U_{2k} , \tag{27a}$$

$$\frac{\partial U_{2k}}{\partial t} = -\Gamma_{22}U_{2k} + \delta_k U_{1k} + \omega_R(t)U_{3k} , \tag{27b}$$

$$\frac{\partial U_{3k}}{\partial t} = -\Gamma_{33}\left(U_{3k} + 1\right) - \omega_R(t)U_{2k} , \tag{27c}$$

with the phenomenological damping constants

$$\Gamma_{11} = \Gamma_{22} = \frac{1}{T_2} , \quad \Gamma_{33} = \frac{1}{T_1} . \tag{28}$$

These equations are the optical Bloch equations, which have been used intensively in the optics of two-level atoms (Slichter, 1992; Allen and Eberly, 1975; Meystre and Sargent, 1990). Note that the inversion relaxes to the full valence band, i.e. $U_{3k} \rightarrow -1$. The Bloch equations can be put into a simple vector equation

$$\frac{\partial}{\partial t}\mathbf{U}_k = -\Gamma \bullet (\mathbf{U}_k + \mathbf{c}_3) + \omega \times \mathbf{U}_k , \tag{29}$$

where the vector of the rotation frequency is

$$\omega = -\omega_R(t)e_1 + \delta_k e_3 , \tag{30}$$

and the diagonal damping matrix

$$\Gamma_{ij} = \Gamma_{ii}\delta_{ij} . \tag{31}$$

From classical mechanics we know that

$$\frac{\partial \mathbf{r}}{\partial t} = \omega \times \mathbf{r} \tag{32}$$

describes the rotation of the vector $\mathbf{r}$ around ω. Disregarding damping for a moment, the rotation alone does not change the length of the Bloch vector. In the ground state the Bloch vector is $U_k = -e_3$, i.e. its length is one. The light field and the detuning cause a rotation of this unit vector. The field-induced rotations around the $-e_1$ axis are called *Rabi flops*. A rectangular pulse of width $\omega_R \Delta t = \pi/2$, e.g., turns the Bloch vector from the ground state through an angle of $\pi/2$ around the $-e_1$ axis $U_k = -e_2$. A $\pi/2$ pulse generates thus from the ground state a maximum polarization. For a finite detuning δ_k, the polarization will then start to rotate around the e_3 axis. The dispersion of the electron energies in the bands automatically introduces a spread of energies. In atomic systems the spread in energies is due to an inhomogeneous broadening.

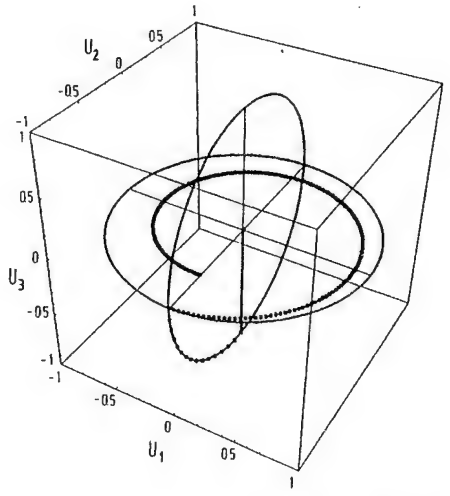


Figure 1. Schematic plot of the rotation of the Bloch vector under a rectangular $\pi/2$ pulse and finite detuning for $T_2 << T_1$.

The facts that a finite pulse can turn the Bloch vector through a certain angle and that the detuning causes a further rotation are the basic ideas for photon echo experiments. Suppose a $\pi/2$ pulse has induced a initial polarization $U_{1k}(0) = 0$, $U_{2k}(0) = -1$. How does it decay after the pulse? Assuming for simplicity that $T_2 << T_1$ the equations with $\omega_R = 0$ which describe this free induction decay are

$$\frac{\partial}{\partial t}U_{2k} = \delta_k U_{1k} - \frac{U_{2k}}{T_2} , \tag{33a}$$

and

$$\frac{\partial}{\partial t}U_{1k} = -\delta_k U_{2k} - \frac{U_{1k}}{T_2} , \tag{33b}$$

with the solutions

$$\begin{bmatrix} U_{1k}(t) \\ U_{2k}(t) \end{bmatrix} = \begin{bmatrix} \cos(\delta_k t) & -\sin(\delta_k t) \\ \sin(\delta_k t) & \cos(\delta_k t) \end{bmatrix} \begin{bmatrix} U_{1k}(0) \\ U_{2k}(0) \end{bmatrix} e^{-it/T_2} . \tag{34}$$

Fig. 1 shows schematically how the polarization spirals for a finite detuning δ_k according to (34) around the z axis after the excitation with a $\pi/2$ pulse. Because of the dispersion $\delta_k = e_k - \omega$, the polarization of electron-hole pairs with different k-values will rotate with different rotation frequencies. If we apply, after the time t_1 a light pulse which causes a rotation of the Bloch vector around the $-e_1$ axis through an angle π, we keep the Bloch vector in the x-y plane. A polarization component which rotated in the time t_1 through the angle α will find itself after the π pulse at the angle $-\alpha$. As all polarization components continue to rotate around e_3 with δ_k they will all return to the origin after the time $2\,t_1$. The coherent superposition of all polarization components will cause the emission of a light pulse, the so-called *photon echo*. Naturally, the intensity of the photon echo will decrease with time as

$$[e^{-2t_1/T_2}]^2 = e^{-4t_1/T_2} .$$

By varying the time delay t_1 between the two pulses one can measure with a photon echo experiment the transverse relaxation time T_2.

INTERBAND QUANTUM KINETICS WITH LO-PHONON SCATTERING

We extend the treatment of the light-induced interband kinetics of the preceding section by taking additionally into account the scattering of the carriers by longitudinal optical (LO) phonons following Haug (1992). In polar semiconductors, this interaction is responsible for the first rapid stage of the cooling of hot carriers on a sub-picosecond time scale. The final stage of the cooling process due to the scattering with acoustic phonons occurs on a nanosecond time scale. The influence of the LO-phonon scattering on the lineshape of emission and absorption spectra of polar semiconductors is correspondingly large.

Before we enter into the technical part we want to discuss first why quantum kinetics instead of the usual semiclassical Boltzmann kinetics is needed in the treatment of the scattering processes. The Boltzmann scattering rate for the density of electrons in the conduction band $\rho_{cc,k}(t)$ e.g., has the following structure

$$\frac{\partial \rho_{cc,k}}{\partial t} = -2\sum_q W_{k,q}\,\rho_{cc,k}(t)\left(1-\rho_{cc,k-q}(t)\right)N_q(t)+\dots\ , \tag{35}$$

where $N_q(t)$ is the mean phonon population of state q. The transition rate $W_{k,q}$ per unit time is given by Fermi's golden rule as

$$W_{k,q} = 2\pi g_q^2 \delta\left(e_{c,k}+\omega_q - e_{c,k-q}\right)\ , \tag{36}$$

for the exlicitly given scattering process from k to k-q under the absorption of a phonon. g_q is the matrix element of the elctron-phonon interaction. First we see that the Boltzmann scattering rate is local in time, i.e. only population factors at time t enter and each of the successive collisions is energy conserving. On a short time scale -- and we are interested in the kinetics in femtosecond regime -- the uncertainity relation $\Delta e \Delta t \geq 1$ shows that the energy conservation is no longer valid. In the framework of quantum kinetics the semiclassical scattering rate will be replaced by

$$\frac{\partial}{\partial t}\rho_{cc,k}(t) = -2\sum_q \int_{-\infty}^{t} dt' K_{k,q}(t,t')\,\rho_{cc,k}(t')\left(1-\rho_{cc,k-q}(t')\right)N_q(t')+\dots\ , \tag{37}$$

where the memory kernel is given by

$$K_{k,q}(t,t') = g_q^2 e^{-\gamma_{e,k}+\gamma_{e,k-q}}(t-t')\cos\left(\left(e_{c,k}+\omega_q-e_{c,k-q}\right)(t-t')\right)\ . \tag{38}$$

Within the considered short time intervals the particles behave as coherent quantummechanical waves. The quantum coherence generates a memory to earlier states of the system, so that the scattering rate becomes non-Markovian. The depth of the memory depends one the decay of the coherence of the particle waves, i.e. on the sum of γ_k and γ_{k-q}, but the memory decays also by interference between the various wave components. For large deviation from the energy conservation the memory kernel oscillates so rapidly that its contribution to the time integral becomes small. Furthermore we see that the Boltzmann limit can be recovered, if we assume a short memory depth and pull the population factors at time t out of the time integral. The remaining integral yields for sufficiently large times a broadened energy δ-function. One can expect that under suitable conditions the coherent oscillations with the dispersionless LO-phonon frequency will be contained in the resulting reduced density matrix elements in the form of quantum beats.

After these intuitive introductory remarks we will describe the formal derivation of the quantum interband kinetics and refer for a more detailed description again to the book of Haug and Jauho (1994). The generalized Kadanoff-Baym equation for the two-band particle propagator

$$G^{<}_{\mu\nu,k}(t,t') = i\left\langle a^{+}_{\nu,k}(t')\,a_{\mu,k}(t)\right\rangle$$

is in the equal-time limit given by

$$\left[i\frac{\partial}{\partial t} - e_{\mu,k} + e_{\nu,k}\right]G^{<}_{\mu\nu,k}(t,t) = -dE(t)\sum_\rho\left[(1-\delta_{\mu\rho})G^{<}_{\rho\nu,k}(t,t) - G^{<}_{\mu\rho,k}(t,t)(1-\delta_{\rho\nu})\right]$$

$$+\sum_\rho \int_{-\infty}^{t} dt'\left[\Sigma^{>}_{\mu\rho,k}(t,t')G^{<}_{\rho\nu,k}(t',t) - \Sigma^{<}_{\mu\rho,k}(t,t')G^{>}_{\rho\nu,k}(t',t)\right.$$
$$\left.-G^{>}_{\mu\rho,k}(t,t')\Sigma^{<}_{\rho\nu,k}(t',t) + G^{<}_{\mu\rho,k}(t,t')\Sigma^{>}_{\rho\nu,k}(t',t)\right] \tag{39}$$

The first two terms describe the time-development of free carriers without any damping under the excitation with a coherent laser pulse as discussed in the preceding chapter. The last two lines describe the influence of scattering processes. For the self-energies we will take into account only the generalized Hartree-Fock contribution due carrier-scattering with a thermal bath of LO-phonons. Under these simplifications the scattering with the LO-phonons is a particularly simple model, mainly because the LO frequency $\omega_q \sim \omega_0$ can be taken to be constant. By considering the phonons as a given thermal bath, we neglect the changes of the population, the dispersion and the interaction matrix element of the phonons. If these effects - which are known to be important in the relaxation of hot carriers (see e.g. Shah, 1989)-- have to be included, an additional equation for the kinetics of the phonons has to be added and the self-energies have to be calculated more accurately. In the generalized Hartree-Fock approximation the carrier self-energies contain in the band index not only diagonal but also off-diagonal particle propagators:

$$\Sigma^{<,>}_{\mu\rho,k}(t_1,t_2) = i\sum_q g_q^2 D_q^{<,>}(t_1,t_2)\,G^{<,>}_{\mu\rho,k-q}(t_1,t_2)\ , \tag{40}$$

where only the unperturbed phonon propagators are considered:

$$D_q^{0<}(t_1,t_2) = -i\sum_{\pm} N_q^{\pm} e^{\pm i\omega_0(t_1-t_2)}\ ,\ D_q^{0>}(t_1,t_2) = D_q^{0<}(t_1,t_2)^*\ , \tag{41}$$

with

$$N_q^{\pm} = N + \frac{1}{2} \pm \frac{1}{2}\ ,\quad N = \frac{1}{\exp(\frac{\omega_0}{k_B T})-1}$$

is the thermal phonon distribution and g_q is the Fröhlich interaction matrix element which is given in terms of the dimensionless polaron constant α by

$$g_q^2 = \alpha\frac{4\pi(\omega_0)^{3/2}}{(2m_r)^{1/2}}q^2 V\ ,\ \alpha = e^2\left(\frac{m_r}{2\omega_0}\right)^{1/2}\left(\frac{1}{\varepsilon_\infty}-\frac{1}{\varepsilon_0}\right). \tag{42}$$

Here m_r is the reduced electron-hole mass. Inserting these expressions into the two-band quantum kinetic equation we find with $i\rho_{\mu\nu k}(t) = G_{\mu\nu k}^{<}(t,t)$:

$$\left[\frac{\partial}{\partial t} + i(e_{\mu,k}-e_{\nu,k})\right]\rho_{\mu\nu,k}(t) = idE(t)\sum_\rho\left[(1-\delta_{\mu\rho})\rho_{\rho\nu,k}(t) - \rho_{\mu\rho,k}(t)(1-\delta_{\rho\nu})\right] + \left.\frac{\partial \rho_{\mu\nu,k}}{\partial t}\right|_{scatt} \tag{43}$$

where

$$\left.\frac{\partial \rho_{\mu\nu,\mathbf{k}}}{\partial t}\right|_{scatt} = -\sum_{q,\rho,\pm} g_q^2 N_q^{\pm} \int_{-\infty}^{t} dt' \Big[e^{\pm i\omega_0(t'-t)} G^{>}_{\mu\rho,\mathbf{k}}(t,t') G^{<}_{\rho\nu,\mathbf{k}}(t',t)$$

$$- e^{\pm i\omega_0(t-t')} G^{<}_{\mu\rho,\mathbf{k}}(t,t') G^{>}_{\rho\nu,\mathbf{k}}(t',t) \Big] - [\mathbf{k} \leftrightarrow \mathbf{k}-\mathbf{q}] \,. \tag{44}$$

In order to reduce the two-time propagators on the r.h.s. of (44) we use the generalized Kadanoff-Baym ansatz of Lipavsky *et al.* (1986) which reads in its extension for two bands

$$-iG^{<,>}_{\mu\nu,\mathbf{k}}(t_1,t_2) = \sum_{\sigma} \Big[G^{r}_{\mu\sigma,\mathbf{k}}(t_1,t_2) G^{<,>}_{\sigma\nu,\mathbf{k}}(t_1,t_2) - G^{<,>}_{\mu\sigma,\mathbf{k}}(t_1,t_2) G^{a}_{\sigma\nu,\mathbf{k}}(t_1,t_2) \Big] \,. \tag{45}$$

The connection between the two-point particle propagators and the reduced density matrix therefore are, according to (45)

$$G^{<}_{\mu\nu,\mathbf{k}}(t_1,t_2) = -\sum_{\sigma} \Big[G^{r}_{\mu\sigma,\mathbf{k}}(t_1,t_2)\, \rho_{\sigma\nu,\mathbf{k}}(t_2) - \rho_{\mu\sigma,\mathbf{k}}(t_1) G^{a}_{\sigma\nu,\mathbf{k}}(t_1,t_2) \Big] \tag{46}$$

and

$$G^{<}_{\mu\nu,\mathbf{k}}(t_1,t_2) = -\sum_{\sigma} \Big[G^{r}_{\mu\sigma,\mathbf{k}}(t_1,t_2) \{\delta_{\sigma\nu} - \rho_{\sigma\nu,\mathbf{k}}(t_2)\} - \{\delta_{\mu\sigma} - \rho_{\mu\sigma,\mathbf{k}}(t_1)\} G^{a}_{\sigma\nu,\mathbf{k}}(t_1,t_2) \Big] \,. \tag{46}$$

In contrast to the usual Kadanoff-Baym ansatz, the relations (45) take the causality correctly into account in the sense that the two-time particle propagator develops from its equal time limit at the earlier time according to the appropriate spectral function. For $t_1 > t_2$, the time development is given by the retarded Green function $G^r(t_1, t_2)$, while for $t_2 > t_1$ it is governed by the advanced function $G^a(t_1, t_2)$. These relations are exact for noninteracting particles. Note that in the equal-time limit, the off-diagonal spectral functions $G^r(t_1, t_2)$ and $G^a t_1, t_2)$ vanish, because the equal-time anticommutators vanish for these functions. Therefore, the equal-time limit of the generalized Kadanoff-Baym ansatz remains exact also in the matrix extension. Naturally this ansatz still contains the two-time retarded and advanced Green functions. The philosophy of the further development is that after the causality has been built-in properly, one can now use relatively simple approximations for the two-time spectral functions. Using the generalized Kadanoff-Baym ansatz in the form (47) one can put the resulting equations in the following form

$$\left.\frac{\partial \rho_{\mu\nu,\mathbf{k}}}{\partial t}\right|_{scatt} = -\sum_{q,\rho\sigma\tau,\pm} g_q^2 \int_{-\infty}^{t} dt' \Big\{ G^{r}_{\mu\sigma,\mathbf{k}-\mathbf{q}}(t,t') G^{a}_{\tau\nu,\mathbf{k}}(t',t) e^{\pm i\omega_0(t-t')}$$

$$\times \Big[N_q^{+} \big(\delta_{\sigma\rho} - \rho_{\sigma\rho,\mathbf{k}}(t')\big) \rho_{\rho\tau,\mathbf{k}}(t') - N_q^{\pm} \rho_{\sigma\rho,\mathbf{k}-\mathbf{q}}(t') \big(\delta_{\rho\tau} - \rho_{\rho\tau,\mathbf{k}}(t')\big) \Big] \Big\}$$

$$- \{\mathbf{k} \leftrightarrow \mathbf{k}-\mathbf{q}\} \,, \tag{48}$$

$$= \sum_{q,\sigma\tau,\pm} g_q^2 \int_{-\infty}^{t} dt' \Big\{ G^{r}_{\mu\sigma,\mathbf{k}-\mathbf{q}}(t,t') G^{a}_{\tau\nu,\mathbf{k}}(t',t) e^{\pm i\omega_0(t-t')}$$

$$\times \Big[N_q^{+} \rho_{\sigma\tau,\mathbf{k}}(t') - N_q^{\pm}\, \rho_{\sigma\tau,\mathbf{k}-\mathbf{q}}(t') \pm \sum_{\rho} \rho_{\sigma\rho,\mathbf{k}-\mathbf{q}}(t') \rho_{\rho\tau,\mathbf{k}}(t') \Big] \Big\}$$

$$- \{\mathbf{k} \leftrightarrow \mathbf{k}-\mathbf{q}\} \,. \tag{49}$$

In (48-49), the non-Markovian structure of the quantum kinetic scattering terms becomes clear. The memory kernel of these equations is given by

$$G^{r}_{\mu\sigma,\mathbf{k}-\mathbf{q}}(t,t') G^{a}_{\tau\nu,\mathbf{k}}(t',t) e^{\pm i\omega_0(t-t')} \,,$$

all density matrices enter only at the earlier time t'. The quantum mechanical correlations are the origin of the memory of the system.

The non-Markovian nature of these equations is not only important in ultra-short pulse spectroscopy, but also in stationary spectroscopy, provided the short-time correlations are involved. One example is the theory of the linear absorption tail -- the so-called Urbach tail -- below the band gap of a semiconductor. In linear spectroscopy, the densities may be approximated by $\rho_{cc,\mathbf{k}} = 0$ and $\rho_{vv,\mathbf{k}} = 1$ so that one is left with the equation for the polarization. The nearly universally observed exponential decay of the absorption with decreasing light frequency cannot be explained using a Markovian damping for the polarization $\rho_{cv,\mathbf{k}}$, because it gives rise to a Lorentzian lineshape with an asymptotic power law $\sim\omega^{-2}$. The description of the exponential Urbach tail needs even in a phenomenological theory a non-Markovian damping term in the form

$$\int_{-\infty}^{t} dt' \sum_{\mathbf{k}'} \gamma_{\mathbf{k}\mathbf{k}'}(t-t')\, \rho_{cv,\mathbf{k}'}(t') \,,$$

as has been noted by Bányai *et al.* (1989) in connection with the nonresonant optical Stark effect in the Urbach tail region. The quantum kinetics of the polarization (43) with (48) provides the formal derivation of such damping terms with memory structure which result in an exponential absorption tail due to phonon-assisted optical transitions, as will be demonstrated later. Because the exponential Urbach tail extends over a broad frequency range, it tests the short-time correlations and explains why quantum kinetics is even necessary in this classical problem of stationary, linear spectroscopy!

In order to proceed further we have to make approximations for the spectral functions $G^{r}_{\mu\sigma,\mathbf{k}-\mathbf{q}}(t,t')$ and $G^{a}_{\tau\nu,\mathbf{k}}(t,t')$. Here we will use only spectral functions in zeroth order in the field but damped by the phonon collision broadening $\gamma_j^{u}(\varepsilon) \approx 1.2\alpha\omega_0$. In this way, a certain self-consistency is build in the spectral functions and guarantee the numerical stability of the quantum kinetic equations. Thus we use for the spectral functions the following Wigner-Weisskopf form

$$G^{r}_{\mu\nu,\mathbf{k}}(t,t') \approx -i\delta_{\mu\nu}\theta(t-t') e^{(-i\varepsilon_{\mu,\mathbf{k}} - \gamma_{\mu}(t-t')} \,, \tag{50}$$

with $G^{r}_{\mu\nu,\mathbf{k}}(t,t') = G^{a}_{\mu\nu,\mathbf{k}}(t',t)^{*}$.

In the following we will use again the rotating wave approximation. Then the equation for the polarization reduces to

$$\left(\frac{\partial}{\partial t} + i\delta_{\mathbf{k}}\right) P_{\mathbf{k}}(t) = -i\frac{\omega_R(t)}{2} \big[\rho_{cc,\mathbf{k}}(t) - \rho_{vv,\mathbf{k}}(t) \big] - P_{\mathbf{k}}(t)\big|_{scatt} \,, \tag{51}$$

where

$$P_k(t) = \rho_{cv,k}(t)e^{i\omega t} \ , \ \delta_k = e_k - \omega \ , \tag{52}$$

The equation for the electron density in the conduction band becomes

$$\frac{\partial \rho_{cc,k}(t)}{\partial t} = -i\frac{\omega_R(t)}{2}\left[P_k(t) - P_k^*(t)\right] - \left.\frac{\partial \rho_{cc,k}(t)}{\partial t}\right|_{scatt} . \tag{53}$$

A corresponding equation holds for the electrons in the valence band. These three equations have to be solved together. The scattering rates are given in (48) or (49) together with (50). For the polarization one gets

$$\left.\frac{\partial P_k(t)}{\partial t}\right|_{scatt} = -\sum_{q,\pm} g_q^2 \int_{-\infty}^{t} dt' \Big\{ P_k(t') \Big[e^{i(\pm\omega_0 - e_{c,k-q} + e_{v,k} + \omega - \gamma_c - \gamma_v)(t-t')} \left[N_q^{-+} \pm \rho_{cc,k-q}(t') \right] + e^{i(i(\pm\omega_0 - e_{c,k} + e_{v,k-q} + \omega) - \gamma_c - \gamma_v)(t-t')} \left[N_q^{-+} \pm \rho_{vv,k-q}(t') \right] \Big] \Big\} - \{ \mathbf{k} \leftrightarrow \mathbf{k} - \mathbf{q} \} \tag{54}$$

The damping terms of the polarization P_k, are determined by the sum of scattering rates of two processes. The first one is the scattering out of the occupied state c,k with $\rho_{cc,k} = 1$ accompanied by the absorption of a phonon. The corresponding probability is $\propto N(1-\rho_{cc,k-q})$. The second damping process is the scattering into the empty state $\rho_{cc,k} = 0$ accompanied by the emission of a phonon with the probability $\propto (N+1)\rho_{cc,k-q}$. The sum is $N(1-\rho_{cc,k-q}) + (N+1)\rho_{cc,k-q} = (N + \rho_{cc,k-q})$. Adding also the processes where emission and absorption are interchanged, one obtains the term $[N_q^{-+} \pm \rho_{cc,k-q}]$, which is just the total scattering rate in the first term of (54). An equivalent contribution comes from the scattering in and out of state v,k which is the second contribution in (54). The corresponding scattering rate for the electron density is

$$\frac{\partial \rho_{cc,k}(t)}{\partial t} = \omega_R(t)\mathrm{Im}\{P_k(t)\} - \left.\frac{\partial \rho_{cc,k}(t)}{\partial t}\right|_{den} - \left.\frac{\partial \rho_{cc,k}(t)}{\partial t}\right|_{pol} , \tag{55}$$

where the density-dependent scattering rate is given by

$$-\left.\frac{\partial \rho_{cc,k}(t)}{\partial t}\right|_{den} = 2\sum_q g_q^2 \int_{-\infty}^{t} dt' e^{-2\gamma_r(t-t')} \Big\{ \cos[(e_{c,k} - e_{c,k-q} + \omega_0)(t-t')] \times \Big[N\left(\rho_{cc,k-q}(t') - \rho_{cc,k}(t')\right) + \rho_{cc,k-q}(t')\left(1 - \rho_{cc,k}(t')\right) \Big] \Big\} - \{ \mathbf{k} \leftrightarrow \mathbf{k} - \mathbf{q} \} \ , \tag{56}$$

The polarization-dependent scattering rate is given by

$$-\left.\frac{\partial \rho_{cc,k}(t)}{\partial t}\right|_{pol} = -2\sum_{q,\pm} g_q^2 \int_{-\infty}^{t} dt' e^{-2\gamma_r(t-t')} \mathrm{Re}\left(\pm e^{i(e_{c,k} - e_{v,k-q} \pm \omega_0)(t-t')} P_k^*(t') P_{k-q}(t') \right) . \tag{57}$$

Finally we will give the same collision terms in the completed collision approximation and take only the resonant terms (neglecting P_{k-q} terms) with $\Gamma = \gamma_c + \gamma_v$:

$$-\left.\frac{\partial P_k(t)}{\partial t}\right|_{scatt} = -\sum_{q,\pm} g_q^2 \Bigg\{ P_k(t) \Bigg[\frac{\Gamma + i\left(\pm\omega_0 - e_{c,k-q} + \omega + e_{v,k}\right)}{\left(\pm\omega_0 - e_{c,k-q} + \omega + e_{v,k}\right)^2 + \Gamma^2} \left[N_q^{-+} \pm \rho_{cc,k-q}(t) \right] + \frac{\Gamma + i\left(\pm\omega_0 - e_{c,k} + \omega + e_{v,k-q}\right)}{\left(\pm\omega_0 - e_{c,k} + \omega + e_{v,k-q}\right)^2 + \Gamma^2} \left[N_q^{-+} \pm \rho_{vv,k-q}(t) \right] \Bigg] \Bigg\} - \{ \mathbf{k} \leftrightarrow \mathbf{k} - \mathbf{q} \} \ , \tag{58}$$

and similarly one can get the Boltzmann limit for the scattering of the densities.

Numerical Studies

We will present now a numerical evaluation of the above derived kinetic equations according to Bányai *et al.* (1992). For convenience, we change from the conduction-valence-band picture to the electron-hole picture with parabolic one-particle energies measured from the edge of their bands

$$\varepsilon_{j,k} = \frac{k^2}{2m_j} = \frac{\mu}{m_j}\varepsilon_k \quad \text{with} \quad j = \{e,h\} \ . \tag{59}$$

The electron-hole pairs are resonantly excited by a classical light field of carrier frequency ω with $\omega - E_g = \Delta > 0$. The envelope of the light field is a Gaussian function in time, i.e. the light pulse is given in terms of a time-dependent Rabi frequency $\omega_R(t) = dE_0(t)$. We are assuming that the excitation is isotropic, i.e. the induced polarization and the generated electron and hole densities depend only on $k = |\mathbf{k}|$. For shortness, we denote

$$\rho_{cc,k}(t) = f_{e,k}(t) \ \text{ and } \ 1 - \rho_{vv,k}(t) = f_{h,k}(t)$$

and

$$p_k(t) = \rho_{cv,k}(t)e^{i(E_g + \varepsilon_k)t} \ . \tag{60}$$

Note, that for the numerical evaluation it is advantageous to eliminate the free-carrier oscillations rather than the optical frequency as done in the general theory, because in this case the oscillations are absent after the pulse. For the assumed isotropic excitation one can use instead of the momentum $\mathbf{q}$ the energy $\varepsilon_{k'} = (\mathbf{k}-\mathbf{q})^2/(2\mu)$ as integration variable. With this transformation the quantum kinetic equations can now be written as

$$\frac{df_{j,k}(t)}{dt} = \omega_R(t)\mathrm{Im}\left[p_k(t)e^{-i\delta_k}\right] - \frac{\alpha\omega_0^2}{\pi\sqrt{\omega_0\varepsilon_k}} \int_0 d\varepsilon_{k'} \int_0 dt'$$

$$\times \mathrm{Re}\Big\{K_j(\varepsilon_k,\varepsilon_{k'},t-t')f_{j,k}(t')\big[1-f_{j,k'}(t')\big]$$

$$-K_{j\neq j}(\varepsilon_{k'},\varepsilon_k,(t-t')p_k(t')p_{k'}(t')e^{i(\varepsilon^k-\varepsilon_{k'})t}\Big\}-\{\mathbf{k}\leftrightarrow\mathbf{k}-\mathbf{q}\}\,, \tag{61}$$

and

$$\frac{dp_k(t)}{dt}=i\frac{\omega_R(t)}{2}e^{i\delta_k t}\big[1-f_{e,k}(t)-f_{v,k}(t)\big]-\frac{\alpha\omega_0^2}{\pi\sqrt{\omega_0\varepsilon_k}}\sum_j\int_0 d\varepsilon_{k'}\int_0 dt'$$

$$\times\Big\{\Big[K_j(\varepsilon_k,\varepsilon_{k'},t-t')\big(1-f_{j,k'}(t')\big)+K_j(\varepsilon_{k'},\varepsilon_k,t-t')f_{j,k'}(t')\Big]p_k(t')$$

$$-\Big[K_j(\varepsilon_{k'},\varepsilon_k,t-t')\big(1-f_{j,k}(t')\big)+K_j(\varepsilon_k,\varepsilon_{k'},(t-t')f_{j,k}(t')\Big]p_{k'}(t')e^{i(\varepsilon_k-\varepsilon_{k'})t}\Big\} \tag{62}$$

where the memory kernel is given by

$$K_j(\varepsilon,\varepsilon',t)=\ln\left|\frac{\sqrt{\varepsilon}+\sqrt{\varepsilon'}}{\sqrt{\varepsilon\varepsilon'}}\right|\big[(N+1)e^{-i\omega_0 t}+Ne^{i\omega_0 t}\big]e^{i(\varepsilon_j-\varepsilon_j')(t-\gamma t}\,. \tag{63}$$

The detuning $\delta_k = E_g + \varepsilon_k - \omega$.

For LO-phonons without dispersion, a Markov limit is achieved, unlike by the broad-band acoustical phonons, only due to the presence of the attenuation factor $exp(-\Gamma t)$ in the definition of the memory kernel (62). If the functions vary slowly on the scale of $1/\Gamma$, they can be taken outside the time integral and the remaining time integral may be extended to infinity. This last approximation leads to energy conservation with an uncertainty Γ. The equations are local in time if also the nonsecular terms $\partial\rho_{cc,k}/\partial t$ are neglected (57). This last step is usually motivated by the oscillating character of these terms (having a supplementary factor $exp[i(\varepsilon-\varepsilon')t]$. However, the argument as such does not hold in a continuous spectrum. We shall solve the equations both with and without these simplifying approximations.

The equations of the previous section are numerically solved for various choices of the physical parameters. Two different numerical methods are used. One coincides with the standard Fredholm iteration, the second uses the fact that the memory kernel $K(t\text{-}t')$ can be written as a sum of factorized kernels.

Factorizing Integral Kernel

The exponential form of our integral kernels allow to write them in the form

$$K(t-t')=\sum_i \kappa_i(t)\kappa_i'(t)\,. \tag{64}$$

The set of differential equations has symbolically the structure

$$\frac{df}{dt}=\int_{-\infty}^{t}dt'\sum_i\kappa_i(t)\kappa_i'(t')F(t')=\sum_i\kappa_i(t)G_i(t) \tag{65}$$

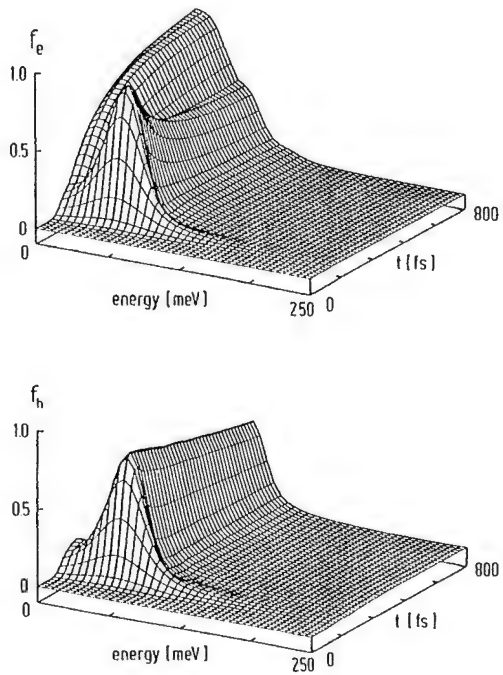


Figure 2. Electron and hole distributions versus energy and time for GaAs excited by a laser pulse.

with the supplementary variables

$$G_i(t)=\int_{-\infty}^{t}dt'\kappa_i'(t')F(t')\,, \tag{66}$$

for which one gets a set of local differential equations

$$\frac{dG_i(t)}{dt}=\kappa_i'(t)F(t)\,. \tag{67}$$

This reformulation transforms the original set of integro-differential equations into a larger set of local differential equations (65) and (66). This method is very rapid and allows to check the convergence of the discretization in time and energy with a high degree of precision. Alternatively a workstation and a vector computer have been used. The following figures have been calculated for GaAs with the parameters $\alpha = 0.069$, $\omega_0 = 36$ meV, $m_h = 0.46\, m_0$, $m_e = 0.067\, m_0$, $\varepsilon_\infty = 11.1$, $\varepsilon_0 = 13.1$ and a lattice temperature $T = 174$ K. For the laser pulse, a Rabi frequency of $\omega_{R,0} = 26.3$ meV, a pulse width of $\tau = 50$ fs, and an excess energy of $\Delta = 60$ meV have been assumed. The pulse is at the central frequency a π-

pulse and induced a maximum pair density of about 10^{18} cm^{-3}. In Fig. 2 and Fig. 3, the energy and time dependencies of the electron and hole populations and of the absolute amount of the polarization are shown. One sees the rapid relaxation within about 200 fs after the pulse. The absolute value of the polarization in Fig. 2 shows in the later stage ($t >$ 200 fs) a camelback structure. At resonance, the pulse corresponds to a π-pulse which creates a maximum inversion and thus a minimum in the absolute amount of the polarization. With increasing detuning away from the resonance the excited inversion decreases and correspondingly the induced polarization increases, until it finally decrease because of the finite spectral width of the pulse. A comparison between the solution of the fully retarded equation and of the completed collision approximation shows only only at low energies minor quantitative differences, but there is no striking qualitative difference between the full quantum kinetics and the nonretarded kinetics of the completed collision approximation with collision broadened energy conserving δ-functions! This has been expected only for times larger than $1/\Gamma$, but actually occurs for all times. The specific features of the coherent dynamics do not show up because of the energy broadening produced by the short pulse. The broad band of excited free-carrier states interfere destructively, so that oscillations of the interband polarization with the phonon frequency do not survive.

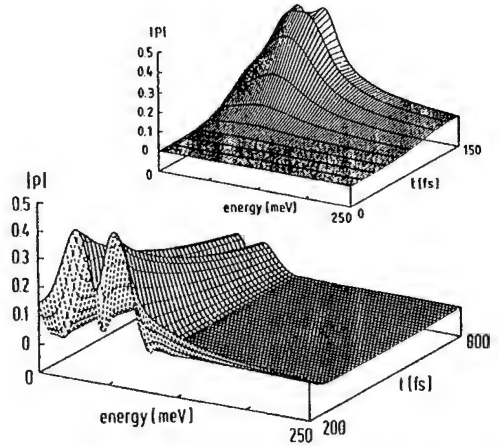


Figure 3. Absolute amount of the induced polarization versus energy and time for GaAs excited by a laser pulse. For better visibility the plot has been split in an early ($t <$ 150 fs) and a late part (200 $< t <$ 800 fs).

In the next section we will demonstrate that if the attractive electron-hole interaction is taken into account, the situation changes. Due to the accumulation of excitonic oscillator strength near the band edge, coherent oscillations of the polarization will become clearly observable in the form of phonon quantum beats.

EXCITON QUANTUM-KINETICS IN POLAR SEMICONDUCTORS

Excitonic effects are known to influence strongly the band-edge optical spectra of semiconductors (see *e.g.* Haug and Koch, 1993). Therefore we will extend our quantum kinetic equations by taking the Coulomb interaction in the Hartree-Fock approximations following Tran and Haug (1993). The main effect of such a treatment is that it contains due to the attractive *e-h* potential automatically the complete exiton spectrum with all bound and ionized states. A similar combination of LO phonon scattering with Coulomb Hartree-Fock terms has been used by Kuhn and Rossi (1993) for the quasi-classical kinetics. Naturally, the Hartree-Fock terms give only a simple mean-field approximation of the carrier interactions but omit any Coulomb scattering. These problems will be addressed later. The Hartree-Fock energy contributes due to the instantaneous Coulomb interaction $V_q(t,t') = V_q\ \delta(t-t')$, with $V_q = 4\pi e^2/(\varepsilon_0\, q^2\, V)$, to the singular part of the self-energy. The contribution to $\Sigma_s(t)$ is

$$\Sigma^{HF}_{\mu\nu,k}(t) = -\sum_q V_q\, \rho_{\mu\nu,k-q}(t)\ . \tag{68}$$

Note that the matrix contains intra- and interband elements. The commutator of Σ^{HF} with the matrix $G^<$ yields the following extension of the coherent Bloch equations:

$$\left[\frac{\partial}{\partial t} - \left(e_{\mu,k} + \Sigma^{HF}_{\mu\mu,k}(t) - e_{\nu,k} - \Sigma^{HF}_{\nu\nu,k}(t)\right)\right] G^<_{\mu\nu,k}(t,t)$$

$$= -dE(t) \sum_\rho \left\{ \left[1-\delta_{\mu\rho}\right] G^<_{\rho\nu}(t,t) - G^<_{\mu\rho}(t,t) \left[1-\delta_{\rho\nu}\right] \right.$$

$$\left. + \left[1-\delta_{\mu\rho}\right] \Sigma^{HF}_{\mu\rho}(t) G^<_{\rho\nu}(t,t) - G^<_{\mu\rho}(t,t) \left[1-\delta_{\rho\nu}\right] \Sigma^{HF}_{\rho\nu}(t) \right\}. \tag{69}$$

One sees already that the intraband Hartree-Fock self-energies give renormalizations of the single-particle energies, while the interband self-energy contributions modify the driving field term, or in other words the Rabi frequency. With the real Coulomb Hartree-Fock self-energies the quantum kinetic equation for the polarization in the electron-hole picture $P_k = \rho_{cv,k}(t)$ is

$$\left[\frac{\partial}{\partial t} + i\delta_k + i\Sigma_k(t)\right] P_k(t) = -i\frac{\omega_{R,k}(t)}{2}\left[1 - f_{e,k}(t) - f_{h,k}(t)\right] - \left.\frac{\partial P_k(t)}{\partial t}\right|_{scat}, \tag{70}$$

where

$$\omega_{R,k}(t) = \omega_R(t) + 2\sum_q V_q\, P_{k-q}(t) \tag{71}$$

is the Rabi frequency renormalized by the action of the attractive *e-h* Coulomb potential V_q acting on the pair function P formulated in k-space. This term can also be seen as a local-field correction of the electric field of the laser light. The *e-h* pair self-energy Σ_k is just the sum of the electron and hole Hartree-Fock or exchange energies

$$\Sigma_k(t) = \sum_{j,q} V_q f_{j,k-q}(t) \text{ with } j = \{e,h\} \ . \tag{72}$$

Note, that the (formally divergent) exchange term with the full valence band does not appear, because it is already included in the original single-particle energies (Schäfer and Treusch, 1986). The corresponding quantum kinetic equation for the electron and hole densities are

$$\frac{\partial f_{j,k}(t)}{\partial t} = \mathrm{Im}\left[\sigma^*_{R,k}(t) P_k(t)\right] - \left.\frac{\partial f_{j,k}(t)}{\partial t}\right|_{den} - \left.\frac{\partial f_{j,k}(t)}{\partial t}\right|_{pol} \ . \tag{73}$$

As long as excitonic effects are not included in the approximations for the spectral functions, the scattering terms in these equations are still given by (50). This approximation however holds only if the exciton binding energy e_0 is small with respect to the phonon energy ω_b.

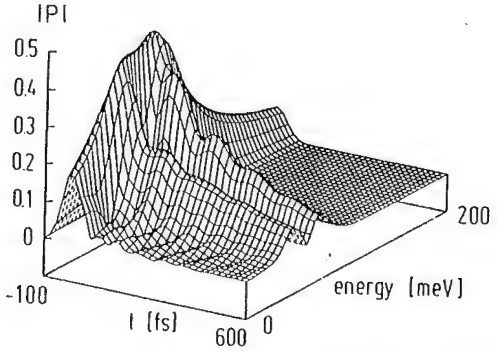


Figure 4. Absolute amount of the interband polarization $P_k(t)$ versus energy $e_{e,k}$ and time for GaAs excited with a 50 fs pulse peaking at $t = 0$ with 60 meV excess energy and a Rabi frequency of $dE_0 = 26.4$ meV.

LO-Phonon Quantum Beats

Quantum beats are the clearest manifestation of optical coherence in the induced polarization, and can e.g. be measured in self-defracting four-wave mixing experiments (for recent reviews see *e.g.* Göbel *et al.*, 1992; Kim *et al.*, 1992). Quantum beats arise when there are transitions from one initial state to two different final states. Quantum beats in exciton systems have e.g. been observed between the light and heavy hole exciton in quantum wells (Leo *et al.*, 1990a; Feuerbacher *et al.*, 1990), between free and bound exciton states (Leo *et al.*, 1990b; Stolz *et al.*, 1991; Pantke *et al.*, 1992), between exciton states split in a magnetic field (Langer *et al.*, 1990), and between excitons from the lower and upper polariton branch (Fröhlich *et al.*, 1992). We propose that a very basic process namely the coupling of the electrons and holes to the LO phonons also gives rise to quantum beats. Actually polarization oscillations with the period of a longitudinal optical phonon mode have been observed (Cho *et al.*, 1990) in transient reflectivity measurements. The excitation of these modes in the experiment was not predominantly due to the intrinisic stimulated Raman-like process which is considered here, but due to a nonlinear process caused by the electron-hole charge separation in the field of surface charges. The quantum beats due to LO-phonons in our model can be seen as the beating between direct and LO-phonon assisted transitions. The same coupling naturally gives also rise to the relaxation of the excited pairs and to the polarization decay, so that this partially coherent process is a typical example for quantum kinetics.

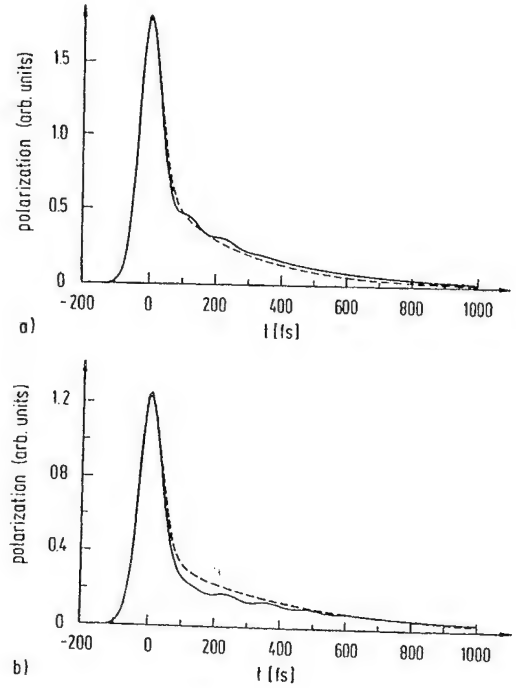


Figure 5. Incoherently summed polarization $P(t) = \Sigma P_k(t)$ versus time for GaAs excited with a 50 fs pulse peaking at $t = 0$ with an excess energy of 100 meV (a) and 20 meV (b) and a Rabi frequency of $dE_0 = 26.4$ meV. Full line: with memory; long-dashed line: without memory

The Gaussian pulses have a width $\tau = 50$ fs so that the spectral width of the excited energy range is larger than the LO-phonon energy. The temperature of the phonons has been assumed to be $T = 300$ K. Figure 4 shows the time and energy dependence of the absolute amount of the polarization $P_k(t)$. At least three oscillation periods with an LO-phonon period can be seen. Figure 5 shows the incoherently summed polarization for an excess energy of 100 and 20 meV, respectively. For excitations deeper in the band the quantum beats become less pronounced, but their amplitudes increase with decreasing excess energy. For 20 meV excess energy one can identify 5 oscillation periods. If also the field-induced off-diagonal matrix elements of the spectral functions are taken into account, the oscillations become even more pronounced (Reitsamer *et al.*, 1994). For nonresonant excitation the differences between quantum kinetics and Markovian kinetics are even more

pronounced. In this exciton regime naturally the theory with Coulomb interaction and without differ very strongly. For resonant excitation, the proposed LO-phonon quantum beats are expected to be smeared out by Coulomb scattering between the excited carriers. However for nonresonant excitation conditions (i.e. with only virtual excitation of electron-hole pairs) one knows e.g. from the optical Stark effect that the influence of Coulomb scattering is negligible. We expect that the proposed LO-phonon quantum beats can best be observed with nonresonant excitation.

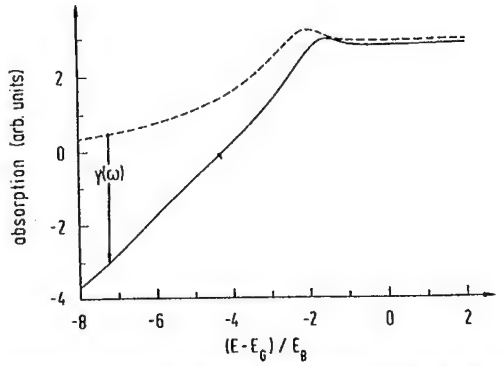


Figure 6. Linear band tail absorption spectrum of GaAs at room temperature. Full line: with memory; long-dashed line: without memory

Urbach Tail Absorption

Finally, we will show that the description of the linear nonresonant absorption for stationary excitation needs the quantum kinetic equations with memory. Naturally under stationary conditions the time integrals can be evaluated explicitly. But the quantum kinetic equations and their Markovian approximation yield different frequency dependencies. The corresponding absorption spectra are shown in Fig. 6 for GaAs. The logarithmic scale shows clearly how quantum kinetics yields over many orders of magnitude an approximately exponential absorption tail, while the Markovian kinetic model has a Lorentzian lineshape. Following the arguments developed by Haug *et al.* (1990), one reaches an exponential lineshape starting from a Lorentzian one only if the linewidth $\gamma = \gamma(\omega)$ decreases rapidly with increasing detuning $\omega_x - \omega$ below the groundstate exciton resonance at ω_x (see arrow in Fig. 6). The decrease of $\gamma(\omega)$ away from the exciton resonance has the simple reason that with increasing detuning $\omega_x - \omega$ it is increasingly difficult to scatter by LO-phonon absorption into real final states with frequency $> \omega_x$. A frequency-dependent damping $\gamma(\omega)$ of the polarization $P(\omega)$ in the time representation of the polarization equation gives rise to a convolution integral $\int dt' \gamma(t-t')P(t')$, i.e. to memory effects. It has been seen above that such memory effects are the trademark of quantum kinetics. Because in the tail absorption relatively large frequency differencies $\omega_x - \omega$ have to be considered, a theory which is correct at correspondingly small times has to be used. This explains why even in the linear, stationary problem of the Urbach tail absorption quantum kinetics is needed. It is known (as discussed *e.g.* by Haug *et al.*, 1990) that in more polar materials vertex corrections of the *e-h* self-energies or even more refined approximations such as cummulant expansions are needed in order to get a quantitative description of the universally observed Urbach absorption tail. But irrespective of such refinements of the theory, the conjecture of Haug *et al.* (1990) that the exponential nature of the absorption tail is a manifestation of the retardations and thus of quantum kinetics is proven here clearly. It should be mentioned that Hartmann and Schäfer (1992) reached the same conclusion from their studies of the quantum kinetic equation of the two-time particle propagators. Furthermore, it is known that at low temperatures LO-phonon sidebands grow out of the Urbach tail of linear absorption. This fact supports our proposal that the calculated LO-phonon quantum beats should be observable for nonresonant sub-bandgap excitation in time-resolved four-wave mixing experiments.

BUILD-UP OF SCREENING

In a dense plasma the scattering between the charged carriers provides the fastest relaxation process. This Coulomb scattering is due to two-particle collisions, the corresponding scattering rate increases therefore roughly with the square of the plasma density, while the LO-phonon scattering increases only linearly. In comparison with the phonon scattering which has been studied so far, the Coulomb scattering has several new features. In order to concentrate on these new properties of Coulomb scattering, we will treat first only the quantum kinetics in a neutral electron gas whose charge is compensated by a positive background, called the jellium model. Coulomb scattering conserves the total momentum and energy of the plasma. Therefore it yields a fast relaxation of the originally excited non-equilibrium plasma into a thermal one, whose temperature is determined by the total energy of the plasma. The relaxation of the plasma temperature to that of the lattice will then take place due to the scattering of electrons with phonons. Furthermore, if we calculate as for the phonon scattering the Boltzmann transition rates by first-order perturbation theory -- called the first Born approximation -- one gets for the long-range Coulomb interaction a divergent result. One overcomes this difficulty by using a partial summation of higher-order diagrams for the effective two-particle Coulomb interaction. In this way the screening of the Coulomb potential by the surrounding carriers is taken into account. In high-density plasmas the random phase theory, abbreviated with RPA, yields a good description of the screening of the Coulomb potential. In equilibrium many-body theory screening is expressed in terms of a momentum- and frequency-dependent dielectric function $\varepsilon(\mathbf{q},\omega)$. The $\mathbf{q},\omega$ dependencies originate from the dependence of the screening on the differences of the space and time coordinates, $\mathbf{r}_1-\mathbf{r}_2$, t_1-t_2 of the two interacting particles. However, in non-equilibrium situations the screened Coulomb potential depends as any other Green function separately on all coordinates $\mathbf{r}_1$, $\mathbf{r}_2$, t_1, t_2. Screening between two newly created carriers, e.g., builds up as the surrounding particles rearrange themselves by scattering. The description of this build-up of screening on a femtosecond time scale needs the dependence of V_s on two times.

Resonant femtosecond pump and probe spectroscopy with semiconductors (Oudar *et al.*, 1985; Knox *et al.*, 1986; Lin *et al.*, 1987; Becker *et al.*, 1988) allows to investigate the regime of ultra-fast relaxation kinetics in an electron-hole plasma governed by Coulomb interactions. Monte Carlo simulations (Osman *et al.*, 1987; Stanton *et al.*, 1988; Goodnick and Lugli, 1988; Bailey *et al.*, 1990; Joshi *et al.*, 1990; El Sayed *et al.*, 1992) and direct numerical integrations (Collet *et al.*, 1983; Collet and Ammand, 1983, 1986; Asche and Sarbei, 1987; Schäfer, 1988; Binder *et al.*, 1992) of the semiclassical Boltzmann equation have been used to describe the time-development of the non-equilibrium electron distributions. As discussed in other contributions, closely related studies of the non-equilibrium electron kinetics are required to describe the transport in semiconductor

microstructures (see also e.g. Jacobini and Reggiani, 1983; Ferry, 1991). The large electric fields in these small devices cause large deviations from equilibrium distributions.

It is obvious that on a very short time scale the Boltzmann picture of individual, successive collisions breaks down and has to be replaced by a quantum kinetic theory. In comparison with the quantum kinetics of electrons with LO-phonon scattering, the quantum kinetics of a dense electron gas with Coulomb interactions is less developed and not so well understood. Following Haug and Ell (1992) we will derive a quantum kinetic description of a dense, spatially homogeneous electron gas with Coulomb interaction in the absence of any field. The idea is to study for this rather basic system the initial quantum kinetic regime, starting from a given initial non-equilibrium distribution (e.g. generated by a femto-second laser pulse). We do not assume that a separation between a "macroscopic" time scale and a "microscopic" time scale is possible. Kadanoff and Baym (1962) used in their pioneering work on a Coulomb system with screening this assumption and showed that it allows a quantum mechanical derivation of the semiclassical Boltzmann equation for this system too.

Our starting point is again the generalized Kadanoff-Baym equation for the equal-time limit of the particle function $G_k^<(t,t) = if_k(t)$, where $f_k(t)$ is the electron distribution function in a single band.

$$\left.\frac{\partial f_k(t)}{\partial t}\right|_{scatt} = \int_{-\infty}^{t} dt' \left[\Sigma_k^>(t,t')G_k^<(t',t) - \Sigma_k^<(t,t')G_k^>(t',t) - G_k^>(t,t')\Sigma_k^<(t',t) + G_k^<(t,t')\Sigma_k^>(t',t)\right] . \tag{74}$$

The initial time is taken to be again $-\infty$ because we treat in all applications the excitation of the carriers by an optical pulse explicitly. In the random phase approximation (RPA) the Coulomb scattering self-energies $\Sigma_k^\lessgtr(t,t')$ are given by

$$\Sigma_k^\lessgtr(t,t') = i\sum_q G_k^\lessgtr(t,t')V_{sq}^\lessgtr(t,t') . \tag{75}$$

The kinetic screened Coulomb potentials $V_{sq}^\lessgtr(t,t')$ will be expressed in terms of the intraband polarization functions $L_q^\lessgtr$ and the spectral functions V_{sq}^r and V_{sq}^a of the screened Coulomb potential.

Relation between $V_{sq}^<(t_1,t_2)$ and $L_{sq}^<(t_1,t_2)$

Using the notation of DuBois (1967) for the non-equilibrium Keldysh Green function matrix, we write (the explicit momentum as well as the time arguments and integrals are suppressed, the time ordering index η is written explicitly assuming the summation convention for equal indices)

$$V_s^{\eta_1\eta_2} = V\delta_{\eta_1\eta_2} + VL^{\eta_1\eta_3}V_s^{\eta_3\eta_2} . \tag{76}$$

Note that V is in this notation a scalar, not a matrix. Particularly one gets

$$V_s^{+-} = V(L^{++}V_s^{+-} + L^{+-}V_s^{--}) . \tag{77}$$

With the relations $A^< = -A^{+-}$, $A^t = A^{++}$, $\tilde{A}^t = -A^{--}$ valid for any Green function, we get

$$V_s^< = V(L^tV_s^< - L^<\tilde{V}_s^t) , \tag{78}$$

where L^t and V^t are time- and antitime-ordered functions. They can be expressed in terms of particle-like, retarded and advanced functions: $L^t = L^t + L^<$ and $\tilde{V}_s^t = V_s^< - V_s^a$. Equation (78) becomes

$$V_s^< = VL^tV_s^< + VL^<V_s^a . \tag{79}$$

The retarded potential obeys the Dyson-like equation

$$V_s^r = V + VL^tV_s^r . \tag{80}$$

Multiplying (80) with the inverse potential V^{-1} from the left, one gets $(V_s^r)^{-1}V_s^r = 1$, with $(V_s^r)^{-1} = V^{-1} - L^r$. Similarly (79) can be written as

$$(V_s^r)^{-1}V_s^< = L^<V_s^a . \tag{81}$$

Multiplying (81) from the left with V_s^r yields the final result

$$V_s^< = V_s^rL^<V_s^a . \tag{82}$$

In extended notation (82) is

$$V_{sq}^\lessgtr(t_1,t_2) = \int_{-\infty}^{t} dt_3 \int_{-\infty}^{t} dt_4 V_{sq}^r(t_1,t_3)L^\lessgtr(t_3,t_4)V_{sq}^a(t_4,t_2) . \tag{83}$$

This result (see also Hartmann *et al.*, 1989) means that the two-time particle-like potential can be expressed exactly in terms of a convolution of the retarded potential, the particle-like polarization $L_q^<$ and the advanced potential. This result is a generalization of the corresponding equilibrium result given e.g. by Kadanoff and Baym (1962). In RPA the polarization bubble is given by

$$L_q^{\eta_1\eta_2}(t_1,t_2) = -2i\eta_1\sum_{k'} G_{k'+q}^{\eta_1\eta_2}(t_1,t_2)G_{k'}^{\eta_2\eta_1}(t_2,t_1) . \tag{84}$$

From (84) one gets

$$L_q^\lessgtr C = -2i\sum_{k'} G_{k'+q}^\lessgtr(t_1,t_2)G_{k'}^{\gtrless}(t_2,t_1) . \tag{85}$$

Inserting these expressions into the scattering rates one finds the form

$$\left.\frac{\partial f_k(t)}{\partial t}\right|_{scatt} = -2\sum_{q,k'}\int_{-\infty}^{t} dt' \int_{-\infty}^{t} dt_1 \int_{-\infty}^{t} dt_2 \Big\{V_{sq}^r(t,t_1)V_{sq}^a(t_2,t') \times \Big[G_{k-q}^>(t,t')G_k^<(t',t)G_{k'+q}^>(t_1,t_2)G_{k'}^<(t_2,t_1) - G_{k-q}^<(t,t')G_k^>(t',t)G_{k'+q}^<(t_1,t_2)G_{k'}^>(t_2,t_1)\Big] - V_{sq}^r(t',t)V_{sq}^a(t_1,t_2)\Big[G_k^>(t,t')G_{k-q}^<(t',t)G_{k'+q}^<(t_2,t_1)G_{k'}^>(t_1,t_2)$$

$$-G^<_k(t,t')G^>_{k-q}(t',t)G^>_{k'+q}(t_2,t_1)G^<_{k'}(t_1,t_2)\Big]\Big\} \ . \qquad (86)$$

One sees that the last two terms are obtained by the first two scattering terms by interchanging k with k-q, k' with k'+q and the retarded potential with the advanced one with reversed time arguments. The last exchange is equivalent to taking the complex conjugate of the spectral functions of the screened potential. Therefore, (86) can be written more concisely as

$$\left.\frac{\partial f_k(t)}{\partial t}\right|_{scatt} = -2\sum_{q,k'}\int_{-\infty}^{t}dt'\int_{-\infty}^{t}dt_1\int_{-\infty}^{t}dt_2\Big\{V^r_{sq}(t,t_1)V^a_{sq}(t_2,t')$$

$$\times\Big[G^>_{k-q}(t,t')G^<_k(t',t)G^>_{k'+q}(t_1,t_2)G^<_{k'}(t_2,t_1)$$

$$-G^<_{k-q}(t,t')G^>_k(t',t)G^>_{k'+q}(t_1,t_2)G^>_{k'}(t_2,t_1)\Big]\Big\}$$

$$-\Big\{k\leftrightarrow k-q,k'\leftrightarrow k'+q,V^{r,a}_{sq}\leftrightarrow V^{r,a^*}_{sq}\Big\} \ . \qquad (87)$$

Before we can use the generalized Kadanoff-Baym ansatz to eliminate the two-time particle functions in terms of the distribution function, we have to establish a definite order between the times t_1 and t_2. By splitting the t_1 integral into an integral from $-\infty$ to t_2 and one from t_2 to t. The resulting kinetic equation takes the form

$$\frac{\partial f_k(t)}{\partial t} = -2\sum_{q,k'}\int_{-\infty}^{t}dt'\Big\{G^r_{k-q}(t,t')G^a_k(t',t)\Big[\Big(1-f_{k-q}(t')\Big)f_k(t')\int_{-\infty}^{t}dt_2V^a_{sq}(t_2,t')$$

$$\times\Big(\int_{-\infty}^{t}dt_1G^a_{k'+q}(t_1,t_2)G^r_{k'}(t_2,t_1)\Big(1-f_{k'+q}(t_1)\Big)f_{k'}(t_1)$$

$$+\int_{-\infty}^{t}dt_1G^r_{k'+q}(t_1,t_2)G^a_{k'}(t_2,t_1)\Big(1-f_{k'+q}(t_2)\Big)f_{k'}(t_2)\Big)V^r_{sq}(t,t_1)\Big]$$

$$-[f\leftrightarrow 1-f]\Big\}-\Big\{k\leftrightarrow k-q,k'\leftrightarrow k'+q,V^{r,a}_{sq}\leftrightarrow V^{r,a^*}_{sq}\Big\} \ . \qquad (88)$$

The collision rates contain the distribution functions of the scattered particles before and after the collision as required by the Pauli principle. However, the initial and final state occupation probabilities of the considered electron which is scattered from $\mathbf{k}\leftrightarrow\mathbf{k}$-$\mathbf{q}$ enter at the retarded time $t'<t$, and those of the scattering partner enter at still earlier times t_1, t_2. These memory effects are the trademark of the quantum kinetic regime. Due to the extra retardation introduced by screening, the quantum kinetics is particularly involved in a dense Coulomb system.

In order to close this quantum kinetic description, one has to calculate the spectral (i.e. retarded and advanced) functions of the screened Coulomb potential. The RPA retarded screened Coulomb potential obeys (80) which is explicitly

$$V^r_{sq}(t_1,t_2) = V_q\delta(t_1-t_2)+V_q\int_{t_2}^{t}dt_3L^r_q(t_1,t_3)V^r_{sq}(t_3,t_2) \ . \qquad (89)$$

The retarded polarization L^r_q can be derived from (84) and reduces with the generalized Kadanoff-Baym ansatz to

$$L^r_q(t_1,t_2) = -2i\sum_{k'}G^r_{k'+q}(t_1,t_2)G^a_{k'}(t_2,t_1)\Big[f_{k'}(t_2)-f_{k'+q}(t_2)\Big] \ . \qquad (90)$$

For the spectral electron functions we use again the Wigner-Weisskopf form

$$G^r_k(t_1,t_2) = -i\theta(t_1-t_2)e^{(-ie_k-\gamma_k)(t_1-t_2)} \ , \qquad (91)$$

where γ_k has to be calculated from the imaginary part of the retarded electron self-energy. But normally a self-consistent treatment of γ_k is omitted and some reasonable collision broadening $\gamma_k\sim\gamma$ will be used.

It is easy to show that (90) reduces in equilibrium to the well-known Lindhard formula. Neglecting for the time beeing the damping in the spectral functions (91), we get with $t=t_1$-t_2 the polarization function

$$L^r_q(t_1,t_2) = -2i\theta(t)\sum_{k}e^{i(e_k-e_{k+q})t}\Big[f_k(t_2)-f_{k+q}(t_2)\Big] \ . \qquad (92)$$

In equilibrium where the distribution functions are time-independent a Fourier transform with respect to t yields the Lindhard formula

$$L^r_q(\omega) = 2\sum_{k}\frac{f_k-f_{k+q}}{\omega+i\delta+e_k-e_{k+q}} \ . \qquad (93)$$

The Coulomb quantum kinetics needs further simplifying approximations in order to make progress with a controlled numerical analysis.

Time-Dependent Plasmon Pole Approximation

A valuable simplification is obtained if we introduce a non-equilibrium version (El Sayed *et al.*, 1994a) of the plasmon pole approximation (see e.g. Haug and Koch, 1993) by considering the long-wavelength limit of the screened potential.

The long-wavelength limit of the polarization function times the Coulomb potential is

$$\lim_{q\to 0}\Big(V_qL^r_q(t_1,t_2)\Big) = -i\lim_{q\to 0}\Big\{\frac{8\pi e^2}{\varepsilon_0q^2}\theta(t_1-t_2)\sum_{k}e^{[i(e_k-e_{k+q})-2\gamma](t_1-t_2)}\Big[f_k(t_2)-f_{k+q}(t_2)\Big]\Big\} \ . (94)$$

With e_k-$e_{k+q}\sim-\mathbf{k}\bullet\mathbf{q}/m$ and $f_{k+q}=f_k+\mathbf{q}\bullet\nabla f_k$ we find the RHS is

$$\lim_{q\to 0}\Big[\frac{8\pi e^2}{\varepsilon_0q^2}\theta(t_1-t_2)\sum_{k}e^{(-iq\bullet k/m-2\gamma)(t_1-t_2)}\mathbf{q}\bullet\nabla f_k(t_2)\Big] \ . \qquad (95)$$

With a partial integration we get finally

$$\lim_{q\to 0}\left(V_q L_q^r(t_1,t_2)\right) = -(t_1-t_2)\theta(t_1-t_2)e^{-2\gamma(t_1-t_2)}\omega_{pl}^2(t_2) \ . \tag{96}$$

This result defines the plasma frequency $\omega_{pl}(t) = (4\pi e^2/\varepsilon_0 m)n(t)$\$ in terms of the total density

$$n(t) = \sum_k f_k(t)$$

of any non-equilibrium distribution. Next we rewrite the Dyson equation (89) for the retarded screened Coulomb potential by introducing a density-density correlation function $S_q(t_1,t_2)$ in the form

$$V_{sq}^r(t_1,t_2) = V_q\left[\delta(t_1-t_2) + S_q(t_1,t_2)e^{-2\gamma(t_1-t_2)}\right] \ . \tag{97}$$

The trivial damping constants γ from the damped free-particle Green functions are taken explicitly into account. For the screening described by the density-density correlation function only the Landau damping contributes. A comparison of (89) and (97) yields

$$S_q(t_1,t_2) = V_q L_q^r(t_1,t_2) + V_q\int_{t_2}^{t_1} dt_3 L_q^r(t_1,t_3)S_q(t_3,t_2) \ . \tag{98}$$

The polarization $L_q^r(t_1,t_3)$ has to be evaluated with $\gamma = 0$. For the long-wavelength limit of (98) we find with (96) the following differential equation:

$$\frac{d^2 S_{q=0}(t_1,t_2)}{dt^2} = -\omega_{pl}^2(t_1)S_{q=0}(t_1,t_2) \ , \tag{99}$$

which shows that the long-wavelength limit density-density correlation function oscillates with the actual plasma frequency which may change parametrically with time t_1 as the plasma density $n(t_1)$ changes. Furthermore, one finds the following initial conditions

$$S_{q=0}(t_2,t_2) = 0 \ , \text{ and } \left.\frac{dS_{q=0}(t_1,t_2)}{dt_1}\right|_{t_1=t_2} = -\omega_{pl}^2(t_2) \ , \tag{100}$$

where we used

$$\lim_{t_1\to t_2}\theta(t_1-t_2) = 1 \ .$$

We solve equation (99) of a parametric oscillator with the ansatz

$$S_{q=0}(t_1,t_2) = \tilde{S}_0(t_1,t_2)\exp\left(-i\int_{t_2}^{t_1} dt_3\omega_{pl}(t_3)\right) . \tag{101}$$

$\tilde{S}_0(t_1,t_2) = s(t_1)$ obeys the following equation of motion in t_1

$$\frac{d^2 s}{dt^2} - i\frac{d\omega_{pl}}{dt}s - 2i\omega_{pl}\frac{ds}{dt} = 0 \ . \tag{102}$$

We assume that the parametric changes of $\omega_{pl}(t_1)$ are sufficiently small in an oscillation period, so that the second-order derivative of s can be neglected. The remaining equation can be solved by separation of variables. We find

$$\frac{ds}{s} = -\frac{1}{2}\frac{d\omega_{pl}}{\omega_{pl}} \ , \tag{103}$$

with the solution

$$s(t_1) = \tilde{S}_0(t_1,t_2) = S_0\omega_{pl}^{-1/2}(t_1)\omega_{pl}^{1/2}(t_2) \ . \tag{104}$$

The first initial condition (100) yields

$$S_{q=0}(t_1,t_2) = -iS_0\omega_{pl}^{-1/2}(t_1)\omega_{pl}^{1/2}(t_2)\sin\left(\int_{t_2}^{t_1} dt_3\omega_{pl}(t_3)\right) . \tag{105}$$

The second initial condition (100) at $t_1 = t_2$ determines S_0. From

$$-iS_0\omega_{pl}(t_2) = -\omega_{pl}^2(t_2) \ , \tag{106}$$

one gets $S_0 = -i\omega_{pl}(t_2)$.

The final long-wavelength limit of the time-dependent density-density correlation is in the plasmon-pole approximation

$$S_{q=0}(t_1,t_2) = -\theta(t_1-t_2)\omega_{pl}^{3/2}(t_2)\omega_{pl}^{-1/2}(t_1)\sin\left(\int_{t_2}^{t_1} dt_3\omega_{pl}(t_3)\right) . \tag{107}$$

Before we can insert $S_q(t_1, t_2)$ into (97) we have to extend it to finite q values. This can be done by comparing the Fourier transform of the equilibrium density-density correlation $S_q^\omega(\omega)$ in the plasmon pole approximation (see e.g. Haug and Schmitt-Rink, 1984) with respect to the relative time coordinate t_1-t_2 with (107). In equilibrium the time-dependent plasmon-pole approximation is (Haug and Ell, 1992)

$$S_q^0(t_1-t_2) = \int_{-\infty}^{\infty}\frac{d\omega}{2\pi}e^{-i\omega(t_1-t_2)}\frac{\omega_{pl}^2}{(\omega+i\delta)^2-\omega_q^2}$$

$$= -\theta(t_1-t_2)\frac{\omega_{pl}^2}{\omega_q}\sin\left(\omega_q(t_1-t_2)\right) \ , \tag{108}$$

with the dispersion of the effective plasmon pole

$$\omega_q^2 = \omega_{pl}^2\left(1+\frac{q^2}{\kappa^2}\right) + Cq^4 \ . \tag{109}$$

The inverse screening length κ can be expressed in a form which can be used also for non-equilibrium distributions:

$$\kappa^2 = \frac{4\pi e^2}{\varepsilon_0} \int_0 de_k \frac{\rho(e_k) f_k}{e_k} \quad , \tag{110}$$

assuming that the (non-equilibrium) distribution is isotropic, i.e. depends only on the energy e_k; $\rho(e_k)$ is the 3d parabolic density of states. C is a numerical constant. The comparison between (107) and (108) shows that at finite q values one has to use the following non-equilibrium density-density correlation

$$S_q(t_1,t_2) = -\theta(t_1-t_2)\frac{\omega_{pl}^2(t_2)}{\omega_q^{1/2}(t_1)\omega_q^{1/2}(t_2)}\sin\left(\int_{t_2}^{t_1} dt_3 \omega_{pl}(t_3)\right) . \tag{111}$$

Naturally the time-dependent frequencies $\omega_{pl}(t)$ and $\omega_q(t)$ have to be calculated from time-dependent $n(t)$, $\kappa(t)$ and $f_k(t)$. With this result the non-equilibrium damped plasmon pole approximation for the retarded screened Coulomb potential (El Sayed *et al.*, 1994a) is obtained with (97) as

$$V_{sq}^r(t_1,t_2) = V_q\left[\delta(t_1-t_2) - \theta(t_1-t_2)\frac{\omega_{pl}^2(t_2)}{\sqrt{\omega_q(t_1)\omega_q(t_2)}}\sin\left(\int_{t_2}^{t_1} dt_3 \omega_{pl}(t_3)\right) e^{-2\gamma_q(t_1-t_2)}\right]. \tag{112}$$

Numerical Studies

We will present numerical studies of the build-up of the screening due to El Sayed *et al.* (1994a) by a straightforward generalization of the above given formula to two kinds of carriers, namely electrons and holes. In particular we will solve the integral equation (98) for the density-density correlation function using the densities $f_{j,k}$ which are obtained from the solution of the coherent semiconductor Bloch equations with the Hartree-Fock Coulomb terms, see (74) to (77) but without scattering terms.

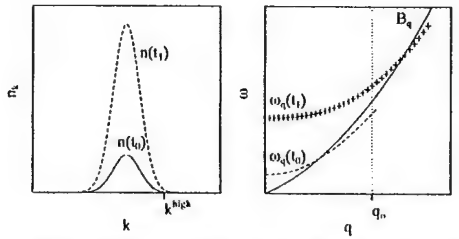


Figure 7. (a)Carrier distribution n_k and (b) plasma frequency ω_q for two times. The full line indicates the boundary of the pair continuum.

Without scattering, $f_{e,k} = f_{h,k}$. These equation will be solved again for a Gaussian pulse with a pulse width of 200 fs and a resonant excitation with a frequency $\omega - E_g = 3E_0$, where $E_0 = e^2/2\varepsilon_0 a_0$ is the exciton Rydberg and a_0 is the exciton Bohr radius. In a two-comonent plasma the time-dependent plasma frequency is in excitonic units

$$\omega_{pl}^2(t) = 16\pi E_0 a_0^3 \sum_j \frac{\mu}{m_j} n_j(t) \quad , \tag{113}$$

where $n_j(t) = \sum_k f_{j,k}(t)$ and μ is the reduced mass. The effective frequency of the plasmon mode (109) becomes

$$\omega_q^2(t) = \omega_{pl}^2(t)\left(1+\frac{q^2}{\kappa^2}\right) + Cq^4 \quad . \tag{114}$$

The screening wavenumber is for isotropic distributions given by

$$\kappa^2(t) = \frac{8}{\pi a_0}\sum_j \frac{m_j}{\mu}\int_0^\infty dk f_{j,k}(t) \quad . \tag{115}$$

The results are again given for bulk GaAs. The amplitude of the pulse is taken to be a π pulse, if the internal field contribution to the Rabi frequency is neglected. The initial value of the density-density correlation is $S_q(t_2, t_2) = 0$. The integral equation (98) defines a recursive procedure in t_1 for any given t_2.

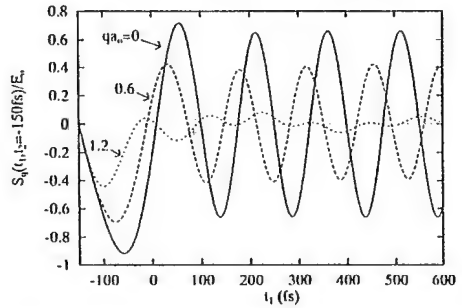


Figure 8. Density-density correlation function versus time t_1 for $t_2 = -150$ fs for various values of qa_0.

Figure 7 shows the density distribution $n_k = f_{e,k} = f_{h,k}$ which are obtained as the solution of the coherent Bloch equations for two times $t_0 < t_1$ and the corresponding plasma frequencies $\omega_q(t)$. The solid line shows the boundary of the pair continuum which is approximately given by $B_q = (2k_{high}q + q^2)/m_e$. Here k_{high} is the momentum of the highest populated states. Beyond the pair continuum the plasma mode suffers a strong Landau damping. For the momentum q_0 in Fig. 7 the plasma frequency is at time t_0 still in the pair continuum, so that the plasma mode is strongly damped. However, at the later time t_1, the frequency has shifted out of the continuum so that the plasma mode can oscillate without Landau damping. Therefore, the boundaries of the onset of Landau damping are smeared out in time. Therefore the calculation of the Landau damping $\gamma_q(t)$ is rather difficult. The resulting density-density correlation function is shown in Fig. 8 as a function of t_1 for a fixed

$t_2 = -150$ fs for three q values. We see pronounced oscillations with $\omega_1(t_1)$ for momentum values between $0 < qa_0 < 1$. We also see that the characteristic time for the build-up of screening is given by $1/\omega_{pl}$ calculated with the density after the pulse. Furthermore, Fig. 8 shows that for $qa_0 > 1$, the plasmon frequency enters the pair continuum and is strongly damped. Similar preliminary results have been obtained by Hartmann *et al.* (1990) for $q = 0$ using equilibrium distributions. In order to investigate in more detail the frequencies contained in the density-density correlation function we want to introduce a Fourier transform which is applicable in non-equilibrium conditions. For slowly varying distributions one defines a Fourier transform of the screened Coulomb potential by

$$V_q(\omega,T) = \int_{-\infty}^{\infty} d\tau\, e^{i\omega\tau} V_q^r\left(T+\frac{\tau}{2}, T-\frac{\tau}{2}\right) , \tag{116}$$

where $T = (t_1 + t_2)/2$ is the central time and $\tau = t_1 - t_2$ is the relative time. However this time integral runs over the past and future, which is no problem in equilibrium situations, but cannot be used in non-equilibrium situations. Instead we propose

$$V_q(\omega,t_1) = \int_{-\infty}^{\infty} dt_2 e^{i\omega(t_1-t_2)} V_q^r(t_1,t_2) = \int_0^{\infty} d\tau\, e^{i\omega\tau} V_q^r(t_1,t_1-\tau) , \tag{117}$$

and correspondingly

$$\varepsilon_q^{-1}(\omega,t_1) = 1 + \int_{-\infty}^{t_1} dt_2 e^{i\omega(t_1-t_2)} S_q(t_1,t_2) . \tag{118}$$

The inverse dielectric function (118) approaches in the limit $t_1 \to -\infty$ the Lindhard formula, but it reflects the correct causal structure, because one integrates only over the past.

The time-dependent dielectric function is evaluated for a $\pi/2$ pulse with a width of τ = 50 fs and $\omega = E_g$. Figure 9 shows the spectra (full lines) of the real and imaginary part of the resulting inverse dielectric functions for three t_1 values, qa_0=1, and γ =0. Up to $t_1 \sim 90$ fs the spectrum is flat, i.e. there is no screening. One plasma oscillation period after the pulse at $t_1 \sim 160$ fs, a broadened plasmon pole emerges. The smaller oscillations around the plasmon resonance are caused by the finite time interval. If one uses a finite γ these oscillations are damped. Finally, we investigate to which $\varepsilon_q^{-1}(\omega,t_1)$ can be described by a plasmon pole approximation developed above. Because the pulse duration is small compared to the period of a plasma oscillation, we can take

$$\omega_{pl}(t_1) = \theta(t_1 - t_0)\omega_{pl}^{final} , \tag{119}$$

where $t_0 = 0$ is the temporal position of the pulse maximum. The resulting plasmon pole approximation for the inverse dielectric function is

$$\varepsilon_q^{-1,PPA}(\omega,t_1) = 1 + \int_0^{t_1} d\tau\, e^{i\omega\tau} S_q^{PPA}(\tau) , \tag{120}$$

with

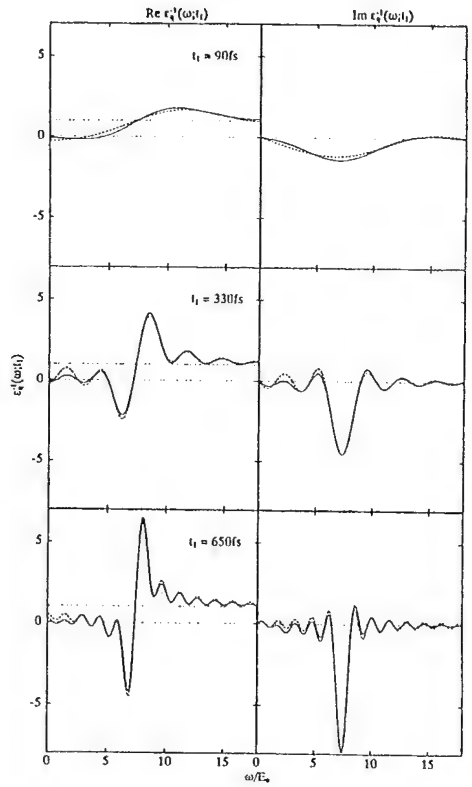


Figure 9. Spectra of the real and imaginary part of $\varepsilon_q^{-1}(\omega,t_1)$ for qa_0 =1 and for three times t_1. Full lines: numerical calculations; dashed lines: time-dependent plasmon pole approximation.

$$S_q^{PPA}(\tau) = -\frac{(\omega_{pl}^{final})^2}{\omega_q}\sin(\omega_q\tau)e^{-\gamma\tau} . \tag{121}$$

Because the intergation in (120) runs only over a finite time interval t_1 the spectrum of the inverse dielectric function broadens due to the uncertainity relation. The dashed lines in Fig. 9 give the plasmon-pole result for $\varepsilon^{-1,PPA}$ at three times t_1 = 90, 330, 650 fs. The following parameters have been used for the best fit of the plasmon pole approximation to the calculated one. At time t_1 =90 fs, we used $\omega_q(t_1) = 6.7E_0$ and $\Gamma(t_1) = 0.8E_0$; at the time t_1 = 330 fs, we used $\omega_q(t_1) = 7.39E_0$ and $\Gamma(t_1) = 0.18E_0$; and at time t_1 =650 fs, we used $\omega_q(t_1) = 7.45E_0$ and $\Gamma(t_1) = 0.15E_0$. The value used for ω_q were taken slightly larger than ω_{pl} in order to account for the dispersion of the plasmon mode. The slight increase in the effective plasma frequency ω_q shows that the average frequency still has a small time dependence and is not completely independent of time as assumed in (119). One sees that even at qa_0 =1 the

plasma mode suffers no Landau damping. We used the damping Γ mainly to model the initial amplitude variation due to the initially increasing plasma frequency. Figure 9 shows that the time-dependent plasmon pole approximation gives even in its simplified form of (121) an excellent description of the femtosecond dynamics of screening.

FEMTOSECOND COULOMB QUANTUM KINETICS AND OPTICAL DEPHASING

The time required for the build-up of screening is approximately given by the inverse plasma frequency. For this reason one can attempt to describe the early phase of a dense pulse-excited electron-hole plasma without any screening. For this purpose we replace the screened retarded and advanced potentials by the nacked Coulomb potential, i. e. we put $V^r_{sq}(t,t') = V^a_{sq}(t,t') = V_q \delta(t-t')$. Such an approximation can only be used in the context of quantum kinetics. In the classical Boltzmann kinetics an unscreened Coulomb potential yields a diverging scattering rate. The absence of a strict energy conservation in the scattering rates of quantum kinetics allows to use this radical approximation in the time interval $\Delta t < \omega_{pl}^{-1}\big|_{max}$, if the peak of the femtosecond laser pulse is at $t = 0$. With this approximation one neglects oscillations of the potential $e^{i\omega_{pl}t}$. Consistent with this approximation is that one also neglects oscillations with the difference of single-particle energies, e.g. terms as $e^{i(\varepsilon_{e,k}-\varepsilon_{e,k-q})t}$. Such terms arise from the spectral functions G^r and G^a.

We switch from the band picture with (c,v) to the electron-hole (e,h)-picture. The quantum kinetic equations are again given by (74) to (77). With the approximation of an instantantaneous unscreened Coulomb potential and with the neglection of oscillation with differences of the particle energies the scattering rates reduce for isotropic distributions to (for a detailed derivation see Haug and Jauho, 1994)

$$\left.\frac{\partial f_{e,k}(t)}{\partial t}\right|_{scatt} = -4\sum_{q,k'j=e,h} V_q^2 \int_{-\infty}^{t} dt' \Big[f_{e,k}(t') - f_{e,k-q}(t') \Big]$$

$$\times \Big[f_{i,k'}(t') \Big(1 - f_{i,k'+q}(t')\Big) - \mathrm{Re}\Big(P_{k'+q}(t') P^*_{k'}(t')\Big) \Big] \,, \qquad (122)$$

and

$$\left.\frac{\partial P_k(t)}{\partial t}\right|_{scatt} = -4\sum_{q,k'j=e,h} V_q^2 \int_{-\infty}^{t} dt' \Big[P_k(t') - P_{k-q}(t') \Big]$$

$$\times \Big[f_{i,k'}(t') \Big(1 - f_{i,k'+q}(t')\Big) - \mathrm{Re}\Big(P_{k'+q}(t') P^*_{k'}(t')\Big) \Big] \,. \qquad (123)$$

We will now report numerical solutions of the quantum kinetic equations (74) to (77) with (122) and (123) according to El Sayed *et al.* (1994b) for a short femtosecond laser pulse. Because one excites with these short pulses states deep in the band where the parabolic band approximation no longer holds we introduce a cut-off wavenumber in the dipole matrix element in the form

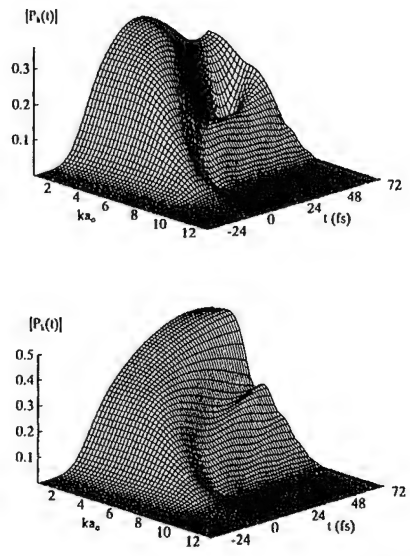


Figure 10. Absolute amount of the optically induced polarization versus time and momentum for a $\pi/8$ pulse with $\tau = 12$ fs and $\Delta\omega = 5E_0$. Upper part: Full quantum kinetics, lower part: coherent Bloch equations without scattering.

$$d_k = d_0 \left[\exp\left(\varepsilon_k - \frac{\varepsilon_c}{\Delta_c} \right) + 1 \right]^{-1} \,. \qquad (124)$$

The cut-off energy $\varepsilon = 64E_0$ has been taken in order to exclude those parts of the band structure which cannot be described in the parabolic approximation. This is particularly important in the local field correction

$$\sum_{k'} V_{|k-k'|} P_{k'}$$

which converges only slowly. The width of the cut-off area has been taken to be $\Delta_c = 0.25E_0$. The total induced density and polarization depend on these cut-off parameters, but the resulting polarization decay and the corresponding decay times are found to be rather insensitive. For the example of GaAs with the parameters $a_0 = 14$ nm and $E_0 = 4.2$ eV, we present results for a Gaussian pulse with a duration of 12 fs and an excess energy $\Delta\omega = \omega - E_g = 5E_0$. The Rabi frequency corresponds to a $\pi/8$ pulse, i.e.,

$$\int_{-\infty}^{\infty} dt d_0 E(t) = \frac{\pi}{8} \,, \qquad (125)$$

which produces a final carrier density of $na_0 = 0.364$ corresponding to $n = 1.32\times10^{17}$ cm^{-3}. The period of a corresponding plasma oscillation is 38 fs. Therefore, we can use the above derived quantum kinetic equation for about the first 40 fs after the pulse maximum. The resulting momentum distributions of the absolute amount of the polarization $|P_k(t)|$ and the distribution $f_{e,k}(t) = f_{h,k}(t)$ are shown in the time interval -24 fs < t < 72 fs in Fig. 10 and Fig. 11. The top part of both figures show the result of the full quantum kinetic equations, the lower part the results of the coherent Bloch equations without scattering terms.

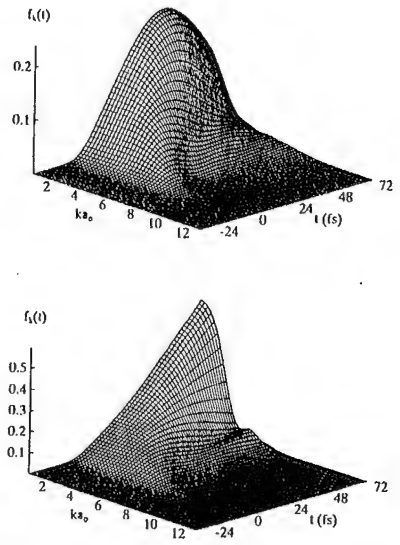


Figure 11. Optically excited carrier distribution versus time and momentum for a $\pi/8$ pulse with τ = 12 fs and $\Delta\omega = 5E_0$. Upper part: Full quantum kinetics, lower part: coherent Bloch equations without scattering.

In the higher k-states one sees in Fig. 10 a weak adiabatic following of the polarization with the pulse. The weak oscillations which are seen clearly in the solution of the coherent Bloch equation can be attributed to beats of the exciton with the continuum states due to the local field correction of the Rabi frequency by the electron-hole attraction. Also the increase of the polarization at small k values after the pulse is due to the local field. In the upper part one sees a decay of the polarization at smaller k values due the quantum scattering kinetics. The successive increase of the polarization at times $t > 0$ fs is outside the validity of our early stage quantum kinetics. Thus the scattering kinetics seems to have most influence in the range $ka_0 < 4$ and 20 fs < t < 40 fs, i.e. well after the pulse. This can be understood further by noting that the coherent Bloch kinetics conserves the unit length of the Bloch vector, as discussed earlier. The length squared is given by

$$B_k(t) = \left(1-2f_k(t)\right)^2 + 4\left|P_k(t)\right|^2 \le 1 \ . \tag{126}$$

Now one notes that the scattering rates of the polarization and the carrier density are in the limit of vanishing q proportional to

$$B_k(t) - 1 = -4\left[f_k(t)\left(1-f_k(t)\right) - \left|P_k(t)\right|^2\right] \ . \tag{127}$$

Because $B_k(t) = 1$ in the coherent regime, the scattering rates are strongly suppressed as long as the dynamics is coherent. This explains the retarded onset of the scattering kinetics well after the pulse. The time development of the incoherently summed polarization $|P(t)|=\Sigma d_k|P_k(t)|$ is shown in Fig. 12. The long-dashed line is the result of the coherent dynamics only, the full line describes the result of the full quantum kinetics, and the short-dashed line represents the the results without the terms proportional to P^2 in the quantum kinetical scattering rates. The polarization decay without the P^2 scattering rates is faster than the decay due to the full kinetics, because the above described "coherent suppression" of the scattering rate is destroyed by the omission of the P^2 terms. While the decay without P^2 terms can be fitted by an exponential decay (crosses in Fig. 12), the decay law of the full Coulomb quantum kinetics is non-exponential. In fact it can be approximated in a certain time window rather well by the result of a quasi-classical theory of the phase decay which has been developed by Gurevich *et al.* (1990). We will describe this rather different, elegant approach to the polarization decay by Coulomb scattering in the next section.

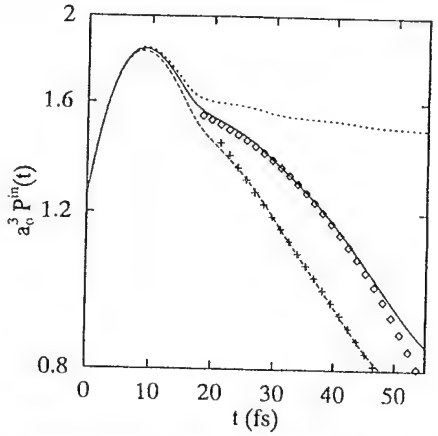


Figure 12. Incoherently summed polarization versus time for a $\pi/8$ pulse with τ = 12 fs and $\Delta = 5E_0$. Full line: Coulomb quantum kinetics, the diamonds: fit with quasi-classical theory; long-dashes: Coherent Bloch equations; short-dashes: Quantum kinetics without P^2 terms, crosses: exponential fit.

While the time intervals in which the phase decay and the carrier redistribution occur are comparable with the corresponding experimental results of Becker *et al.* (1988) and of Elsässer *et al.* (1991), the pronounced non-exponential decay of the polarization has so far not been detected experimentally. One reason for this failure may be, that the experimental results with short femtosecond pulses may also be influenced by the LO-phonon scattering, which we have already discussed in earlier chapters. Particularly for the rather hot carriers exited by the necessarily spectrally broad pulses the LO-phonon scattering rates become comparable to the Coulomb scattering rates.

Quasi-Classical Theory of the Polarization Decay

A rather different description of the polarization decay has been proposed by Gurevich *et al.* (1990), which - however - will turn out to be closely related to the early stage of the Coulomb quantum kinetics with a bare Coulomb potential and without energy conservation. The quasi-classical theory is used to calculate the femtosecond decay of an initially given interband polarization. We present here an adoption of the theory of Gurevich *et al.* (1990) by El Sayed *et al.* (1994b), in which we consider in more detail the nature of a pair excitation in a semiconductor.

In this theory one treats the motion of a newly created pair of carriers in the electrostatic field of the statistically distributed charges of an already created electron-hole plasma. The non-diagonal element of the reduced density-matrix $\rho_{cv,k}(t)$ is in a two-particle space nothing but the momentum representation of the electron-hole pair wavefunction ψ. We assume that the initial value $\rho_{cv,k}(t=0)$ is somehow prepared by the pulse (this process is not treated) and is considered to be given. Because the quasi-classical theory is an expansion in powers of the Planck's constant $\hbar$, we write it here explicitly. In terms of an action function S we write the pair wavefunction as

$$\psi(\mathbf{x}_e,\mathbf{x}_h) = \exp(\mathrm{i}\frac{S}{\hbar}) \ . \tag{128}$$

From the two-particle Schrödinger equation we find the following equation for the action

$$-\frac{\partial S}{\partial t} = \sum_{j=eh} \frac{(\nabla S)^2}{2m_j} + \sum_{jn} V_{jn} + V_{eh} + E_g - \sum_j \mathrm{i}\hbar \frac{\nabla_j^2 S}{2m_j} \ . \tag{129}$$

Here

$$V_{j,n} = \frac{e_j e_n}{\varepsilon_0 |\mathbf{x}_e - \mathbf{x}_h|} \tag{130}$$

is the bare Coulomb interaction between the carrier j and the $2N$ randomly distributed charges $e_n = \pm e$ of the already existing neutral plasma. Only the high-frequency components of the crystal polarization will contribute to the background dielectric function ε_0 of the crystal. Again screening by the plasma is not considered. In the quasi-classical approximation we neglect the term linear in $\hbar$ and get the time-dependent Hamilton-Jacobi equation

$$-\frac{\partial S_{cl}}{\partial t} = \sum_{j=e,h} \frac{(\nabla S)^2}{2m_j} + \sum_{j,n} V_{j,n} + V_{eh} + E_g \ . \tag{131}$$

For short times after the creation of the pair with $\mathbf{p}_e = \hbar\mathbf{k}$, $\mathbf{p}_h = -\hbar\mathbf{k}$ at time $t = 0$ the classical action S_{cl} is still close to that of the free motion

$$S_{cl} = S_{cl}^0 + \delta S \ , \tag{132}$$

with

$$S_{cl}^0 = \sum_j \left[\mathbf{p}_j \bullet \mathbf{x}_j - \frac{p_j^2}{2m_j} \right] - E_g \ . \tag{133}$$

The linearized equation for δS is

$$\left(\frac{\partial}{t} - \sum_{j=e,h} \frac{\mathbf{p}_j \bullet \nabla_j}{2m_j} \right) \delta S = \sum_{j,n} V_{j,n} + V_{eh} + E_g \ , \tag{134}$$

with the solution

$$\delta S = -\int_0^t dt' U(\mathbf{x}_e(t'), \mathbf{x}_h(t')) \ , \tag{135}$$

where

$$\mathbf{x}_j(t) = \mathbf{x}_j + \frac{\mathbf{p}_j t}{m_j} \ . \tag{137}$$

Thus the pair wavefunction is

$$\psi(\mathbf{x}_e,\mathbf{x}_h,t) = \exp\left\{ \frac{\mathrm{i}}{\hbar} \left[\sum_j (\mathbf{p}_j \bullet \mathbf{v}_j - \frac{p_j^2}{2m_j} t) - \int_0^t dt' U(\mathbf{x}_e(t'),\mathbf{x}_h(t')) \right] \right\} \ . \tag{138}$$

The polarization due to an *e-h* pair created at the same position with opposite momenta is given by $\langle a_{h,-k} a_{e,k} \rangle$. Within the quasi-classical approximation we associate with this polarization the pair wavefunction with $\mathbf{x}_e = \mathbf{x}_h$ and opposite momenta. We put the initial *e-h* coordinate equal to zero $\mathbf{x}_e = \mathbf{x}_h = 0$, because the polarization in a spatially homogeneous system is independent of the position. A divergent Coulomb self-energy term in the phase has to be compensated, but does not play a physical role. Now the pair wavefunction has to be averaged over all configurations of the frozen charges of the plasma. The Coulomb interaction with the charges of the plasma is the only term is influenced by the averaging. Thus one has to calculate

$$\left\langle \exp\left\{ -\frac{\mathrm{i}}{\hbar} \int_0^t dt' \sum_{jn} V_{jn}(t') \right\} \right\rangle \ , \tag{139}$$

of the electron or hole which have been at $t = 0$ at the origin, i.e. $\mathbf{x}_j(t') = \mathbf{p}_j t'/m_j$. The averaging is thus

$$\langle \ldots \rangle = \sum_{e_n = \pm e} \prod_{n=1}^{2N} \int \frac{d^3 x_n}{2V} \exp\left\{ -\frac{\mathrm{i}}{\hbar} \int_0^t dt' \sum_j V_{jn}(t') \right\}$$

$$= \left\{ 1 + \frac{1}{N} \sum_{e_n = \pm e} \int \frac{d^3 x_n}{2V} \left[\exp\left(-\frac{\mathrm{i}}{\hbar} \int_0^t dt' \sum_j V_{jn}(t') \right) \right] - 1 \right\}^N \ . \tag{140}$$

Following Chanderasekhar (1943), we use the relation

$$\lim_{N\to\infty}\left(1+\frac{x}{N}\right)^N = e^x$$

and get in the thermodynamic limit with $N \to \infty$, $V \to$ and $N/V = n$ finite, the exponential form

$$\exp\{-n\int d^3x\left[1-\cos\left(\frac{e^2}{\varepsilon_0\hbar}\int_0^t dt'\left(\frac{1}{\left|\frac{\hbar k t'}{m_c}-x\right|}-\frac{1}{\left|\frac{\hbar k t'}{m_h}+x\right|}\right)\right)\right]\}\ , \qquad (141)$$

where the cosine has been generated by the sum over both kinds of charges in the plasma. Using the excitonic units a_0 and $E_0 = e^2/(2\varepsilon_0 a_0) = \hbar^2/(2\mu a_0^2)$ with the reduced electron-hole mass μ, one can make all integration variables dimensionless. The rescaled result is

$$\exp\left\{-na_0^3\left(\frac{2E_0 t}{\hbar}\right)^3(ka_0)^3 2\pi\int_0^\infty r^2 dr\int_{-1}^{1} dz\left[1-\cos\left(\frac{1}{ka_0}\int_0^{1/r} du\right.\right.\right.$$

$$\left.\left.\left.\times\left(\frac{1}{\sqrt{1-2z\alpha_e u+\alpha_e^2 u^2}}-\frac{1}{\sqrt{1+2z\alpha_h u+\alpha_h^2 u^2}}\right)\right)\right]\right\}\ , \qquad (142)$$

where $\alpha_j = \mu/m_j$ is the the mass ratio. This result yields in a d-dimensional system the unusual decay law

$$|P_k(t)| = |P_k(0)|e^{-c_d n a_0^d (ka_0)^d t^d}\ , \qquad (143)$$

with a constant c_d. The time argument t^d in the exponent originates through the rescaling from the d-dimensional volume integral. In a three-dimensional system, (143) describes an extraordinary fast decay with a t^3 in the exponent, in a two-dimensional quantum well structure one gets a Gaussian decay. In both cases the exponent is proportional to the d-dimensional plasma density $n = N/L^d$. For small momenta k the decay becomes increasingly slow, however in this limit the quasi-classical theory looses its validity. The derivation of the quasi-classical phase decay made it clear that as in our previously described femtosecond Coulomb quantum kinetics screening and energy conservation do not enter the theory.

The result of the quasi-classical theory can be fitted to the results of our Coulomb quantum kinetics quite well, as is shown by the diamonds in Fig. 12. Because the excitation of the polarization by the light pulse is not contained in Gurevich's theory, the two descriptions can only be compared in the approximate time interval 20 fs $\le t \le$ 40 fs. If we write the polarization decay due to the quasi-classical theory as

$$|P_k(t)| = |P_k(0)|\exp\left(-\frac{(t-t_0)^3}{t_G^3}\right)\ , \qquad (144)$$

the best fit shown in Fig. 12 has been obtained with $t_0 = -12$ fs and a relaxation time of $t_G =$ 73 fs. One sees that one has to take into account that alrady the first part of the pulse with the center at t_0 prepares the coherent polarization. One also finds that the Coulomb quantum kinetics reproduces the quasi-classical $n^{-1/3}$ dependence of the relaxation time t_G for bulk

REFERENCES

Allen, L., and Eberly, J. H., 1975, "Optical Resonance and Two-Level Atoms," Wiley, New York, reprinted (1987) with corrections, by Dover, New York

Asche, M., and Sarbei, O. G., 1987, *Phys. Stat. Sol B* 141:487.

Ashcroft, N.W., and Mermin, N. D., 1976, "Solid State Physics," Saunders College (HRW), Philadelphia.

Axt, V. M., and Stahl, A., 1993, *Z. Phys. B* to be published.

Binder, R., Scott, D., Paul, A.E., Lindberg, M., Henneberger, K., and Koch, S.W., 1992, *Phys. Rev. B* 45:1107.

Bailey, D.W., Artaki, M. A., Stanton, C. J., and Hess, K., 1987, *J. Appl. Phys.* 62:4638.

Bailey, D.W., Stanton, C. J., and Hess, K., 1990, *Phys. Rev. B* 42:3423.

Bányai, L., Wicht, T., and Haug, H., 1989, *Z. Phys. B* 77:343.

Bányai, L., Tran Thoai, D. B., Remling, R., and Haug, H., 1992, *Phys. Stat. Sol. B* 173:149.

Becker, P.C., Fragnito, H., Brito-Cruz, C., Shaw, J., Fork, R. L., Cunningham, J. E., Henry, J. E., and Shank, C.V., 1988, *Phys. Rev. Lett.* 61:1647.

Binder, K., and Kalos, M.H., 1988, *in* "Monte Carlo Methods in Statistical Physics," Binder, K., ed., Springer, New York, Berlin.

Callaway, J., 1974, "Quantum Theory of the Solid State," Part A, Academic, New York.

Cercinani, C., 1988, "The Boltzmann Equation and its Applications," Springer, New York, Berlin.

Chanderasekhar, S., 1943, *Rev. Mod. Phys.* 15:1.

Cho, G.,C., Kütt, W., and Kurz, H., 1990, *Phys. Rev. Lett.* 65:764.

Collet, J., Ammand, T., and Pugnet, M., 1983, *Phys. Lett.* 96A:368.

Collet, J., and Ammand, T., 1986, *J. Phys. Chem. Sol.* 47:153.

Danielewicz, P., 1984, "Quantum Theory of Nonequilibrium Processes," I and II., *Ann. Phys.* (N.Y.) 152:239,305.

DuBois, D. F., 1967, *in* "Lectures in Theoretical Physics IX C, Kinetic Theory," Brittin, W. E., Barut, A. O., and Guenin, M., eds., Gordon and Breach, New York ,p. 469.

Elsässer, T., Shah, J., Rota, L., and Lugli, P., 1991, *Phys. Rev. Lett.* 66:1757.

El Sayed, K., Schuster, S., Haug, H., Herzel, F., and Henneberger, K., 1994a, *Phys. Rev. B* 49:7337.

El Sayed, K., Bányai, L., and Haug, H., 1994b, *Phys. Rev. B*, to be published.

Ferry, D. K., 1991, "Semiconductors," Mac Millan, N.Y.

Feuerbacher, B. F., Kuhl, J., Eccleston, R., and Ploog, K., 1990, *Sol. State Commun.* 74:1279.

Fröhlich, D., Kulik, A., Uebbing, B., Langer, V., Stolz, H., and von der Osten, W., 1992, *Phys. Stat. Sol. B* 173:31.

Göbel, E. O., Koch, M., Feldmann J., von Plesen, G., Meier, T., Schulze, A., Thomas, P., Schmitt-Rink, S., Köhler, K., and Ploog, K., 1992, *Phys. Stat. Sol. B* 173:21.

Goodnick, S. M., and Lugli, P., 1988, *Phys. Rev. B* 38:10135.

Gurevich, V. L., Muradov, M. I., and Parshin, D.A., 1990, *Europhys. Lett.* 12:375.

Haake, F., 1973, "Springer Tracts Mod. Phys." Springer, New York, Berlin, 66:98.

Hartmann, M., and Schäfer, W., 1992, *Phys. Stat. Sol. B* 173:165.

Haug, H., 1985, *J. Luminesc.* 30:171.

Haug, H., 1988, *in* "Optical Nonlinearities and Instabilities in Semiconductors," H. Haug, ed., Academic, New York, p. 53.

Haug, H., and Henneberger, K., 1991, *Z. Physik B* 84:81.

Haug, H., 1992, *Phys. Stat. Sol. B* 173:139.

Haug, H., Bányai, L., Liebler, J., and Wicht, T., 1990, *Phys. Stat. Sol. B* 159:309.

Haug, H., and Ell, C., 1992, *Phys. Rev. B* 46:2126.
Haug, H., and Koch, S. W., 1993, "Quantum Theory of the Optical and Electronic Properties of Semiconductors," World Scientific, Singapore (2nd ed.).
Haug, H., and Jauho, A. P., 1994, "Quantum Kinetics for Transport and Optics in Semiconductors," Springer, New York, Berlin.
Haug, H., and Schmitt-Rink, S., 1984, *Prog. Quantum Electronics* 9:3.
Hartmann, M., Stolz, H., and Zimmermann, R., 1989, *Phys. Stat. Sol. B* 159:35.
Hartmann, M., and Schäfer, W., 1992, *Phys. Stat. Sol. B* 173:165.
Jacoboni, C., and Reggiani, L., 1983, *Rev. Mod. Phys.* 55:645.
Joshi, R. P., Grondin, R. O., and Ferry, D. K., 1990, *Phys. Rev. B* 42:5685.
Kadanoff, L. P., and Baym, G., 1962, "Quantum Statistical Mechanics," Benjamin, New York.
Kim, D. S., Shaw, J., Schäfer, W., and Schmitt-Rink, S., 1992, *Phys. Stat. Sol. B* 173:11.
Kittel, C., 1967, "Quantum Theory of Solids," Wiley, New York.
Knox, W. H., Hirlimann, C., Miller, D. A. B., Shah, J., Chemla, D. S., and Shank, C. V., 1986, *Phys. Rev. Lett.* 56:1191.
Kuhn, T., and Rossi, F., 1992, *Phys. Rev. Lett.* 69:977.
Kusnetsov, V. A., 1991, *Phys. Rev. B* 44:8721, 13381.
Landau, L. D., and Lifschitz, E. M., 1980, "Statistical Physics," Pergamon, Oxford.
Landau, L. D., and Lifshitz, E. M., 1983, "Theoretical Physics, Vol. X, Kinetics," Pergamon, Oxford.
Langer, V., Stolz, H., and von der Osten, W., 1990, *Phys. Rev. Lett.* 64:854.
Leo, K., Damen, T. C., Shah, J., Göbel, E. O., and Köhler, K., 1990a, *Appl. Phys. Lett.* 57:19.
Leo, K., Damen, T. C., Shah, J., and Köhler, K., 1990b, *Phys. Rev. B* 42:11359.
Lifschitz, E. M., and Pitaevskii, 1981, "Physical Kinetics," Pergamon, Oxford.
Lin, W. Z., Fujimoto, J. G., Ippen, E. P., and Logan, R. A., 1987, *Appl. Phys. Lett.* 50:124
Lindberg, M., Binder, R., and Koch, S. W., 1992, *Phys. Rev. A* 45:1865.
Lipavsky, P., Spicka, V., and Velicky, B., 1986, *Phys. Rev. B* 34:6933.
Lyo, S. K., 1991, *Phys. Rev. B* 43:7091.
Mahan, G. D., 1981, "Many-Particle Physics," Plenum, New York.
Mahan, G. D., 1987, *Physics Reports* 145:253.
Meystre, P., and Sargent III, M., 1990, "Elements of Quantum Optics," Springer, Berlin, Heidelberg.
Mukamel, S., 1990, *Ann. Rev. Phys. Chem.* 41:647.
Osman, M. A., and Ferry, D. K., 1987, *Phys. Rev. B* 36:6018.
Oudar, J. L., Hulin, D., Migus, A., Antonetti, A., and Alexandre, F., 1985, *Phys. Rev. Lett.* 55:2074.
Pantke, K. H., Lyssenko, V. G., Razbirin, B. S., and Hvam, J. M., 1992, *Phys. Stat. Sol. B* 173:69.
Reggiani, L. and Lugli, P. 1988, *J. Appl. Physics* 64:3072.
Schäfer, W., and Treusch, J., 1986, *Z. Phys. B* 63:407.
Schäfer, W., 1988, "Festkörperprobleme (Advances in Solid State Physics)," Vieweg, Braunschweig, 28:63.
Schilp, J., Kuhn, T., and Mahler, G., 1993, *Semicond. Sci. Technol.*, to be published.
Schlösser, J., Neumann, C. H., and Stahl, A., 1992, *J. Phys. Cond. Matter* 4:121.
Scott, D. C., Binder, R., and Koch, S. W., 1992, *Phys. Rev. Lett.* 69:347.
Shah, J., 1989, *Sol. State Electronics* 32:1051.
Stanton, C. J., Bailey, D. W., and Hess, K., 1988, *IEEE J. Quantum Electron.* 24:1614.
Stahl, A., and Balslev, I., 1987, "Electrodynamics of the Semiconductor Band Edge," Springer Tracts in Modern Physics 110 (Springer, Berlin Heidelberg).
Stolz, H., Langer, V., Schreiber, E., Permogorov, S., and von der Osten, W., 1991, *Phys. Rev. Lett.* 67:679.
Tran Thoai, D. B., and Haug, H., 1992, *Phys. Stat. Sol. B* 173:159.
Tran Thoai, D. B., and Haug, H., 1993, *Phys. Rev. B* 47:3574.
Zimmermann, R., 1990, *Phys. Stat. Sol. B* 159:317.
Zimmermann, R., 1992, *J. Luminesc.* 53:187.

CONDUCTANCE IN QUANTUM BOXES: INTERFERENCE AND SINGLE ELECTRON EFFECTS

A. S. Dzurak, M. Field, J. E. F. Frost, I. M. Castleton, C. G. Smith, C.-T. Liang, M. Pepper, D. A. Ritchie, E. H. Linfield, and G. A. C. Jones.

Cavendish Laboratory
Madingley Road
Cambridge, U.K. CB3 0HE

INTERFERENCE IN ELECTRON TRANSPORT THROUGH QUANTUM BOXES

Ballistic electron transport in the two-dimensional electron gas (2DEG) formed at a GaAs/$Al_xGa_{1-x}As$ heterojunction has been studied in depth both experimentally and theoretically in recent years, and remains an active research area. One-dimensional (1D) conductance quantisation (Wharam *et al.*, 1988; van Wees *et al.*, 1988) in a short and narrow constriction provides evidence of ballistic transport and of quantised energy levels due to lateral confinement in the constriction. Models of the constriction which assume electron phase coherence and have a hard-wall lateral confining potential lead to the prediction of length resonance effects in simple split gate devices due to multiple reflections at the ends of the channel (Kirczenow, 1988). These predictions have not proved to be easy to verify experimentally (Smith *et al.*, 1989; Brown *et al.*, 1989a, 1989b; Kouwenhouven *et al.*, 1990; van Wees *et al.*, 1991).

In this section we briefly describe split-gate models in the literature and then justify the use of our simple 1D model. Finally we compare low temperature conductance measurements with numerical calculations for single and double quantum box devices.

The analytical model of a short split-gate device as a quadratic saddle-point, incorporating the smooth nature of the electrostatically defined potential, was developed by Büttiker (1990). This model does not predict length-resonances and may be used to calculate accurately the channel conductance as a function of dc source-drain bias (Patel *et al.*, 1991), temperature (Frost *et al.*, 1993), or with minor modification, number of occupied 1D subbands (Frost *et al.*, 1994), and in the presence of an impurity (Levinson et al., 1992).

Other models of a 1D constriction comprise a channel between two semi-infinite 2DEG planes with a hard wall confining potential (see references 31-60 in van Wees *et al.*, 1991). The channel may be assumed to have abrupt changes in width (Kirczenow, 1988), a linear decrease in width of channel approaching the constriction (Szafer *et al.*, 1989), or a finite radius of curvature to all the corners in the device (Ji, 1993). When sharp corners or hard walls are present, length resonances are predicted due to interference between phase coherent components of the electron wave function reflected from the corners or rapid changes in channel width. These have previously been observed only weakly in the best cases in a simple split gate device (Smith *et al.*, 1989; Brown *et al.*, 1989a, 1989b; Kouwenhouven *et al.*, 1990; van Wees *et al.*, 1991) because it is difficult to pattern the 2DEG on a length scale comparable with the electron Fermi wavelength (Kumar *et al.*, 1989) and have a phase coherence length greater than the device length.

We use a simple 1D barrier transmission probability model without including Coulomb charging effects. Electron phase coherence is implicitly assumed in this model, as in those of the hard-wall models of reference. Either side of the device the 2D potential varies smoothly and we have modelled it by a constant slope, varying between 0 and 9.5 meV over a distance of 1 μm. We can estimate the potential modulation from the 2DEG depth, b, of 90 nm and characteristic period, a, of 200 nm. The potential modulation at the depth of the 2DEG in a periodic structure is given by (Kotthaus et al., 1982)

$$V_{2DEG} = V_s \exp(-2\pi a/b) , \qquad (1)$$

where a typical applied voltage V_s of -1 V at the surface leads to an amplitude of 1.3 mV at the 2DEG.

The conductance G of the device is given by

$$G = \frac{2e^2}{h} \sum_n T_n , \qquad (2)$$

where T_n is the total calculated transmission probability of the device for the nth 1D subband at a particular effective incident energy E (Landauer, 1957, 1970). The total transmission probability is calculated by breaking the potential barrier into thin strips and matching boundary conditions at each interface (see, for example, French and Taylor, 1978).

The split-gate device formation is by electron-beam lithography with Nichrome/gold metallisation, the radius of curvature of the Schottky gate metallisation is greater than 10 nm, and the depth of the 2DEG is 90 nm. The mobility is 9.1×10^5 cm^2/Vs and the sheet carrier concentration is 3×10^{11} cm^{-2} at 4.2 K. The design is based on a unit-cell with a narrow-wide-narrow geometry. The minimum channel width is 300 nm and the period is 200 nm: the device geometries are shown in the inset of Fig. 1. Two-terminal conductance measurements were made at 35 mK with 10 μV , at 300 mK with 30 μV, and at 4.2 K with 100 μV ac excitation using standard phase-sensitive techniques.

Patterning of the underlying 2DEG most closely mirrors the shape of the lithographically defined metallisation at the gate voltage when depletion just occurs under the thin 100 nm width fingers, as shown by a three-dimensional Poisson-Schrödinger calculation (Kumar, 1992). This is at a higher voltage than that needed to deplete carriers from beneath the wider metallisation due to fringing effects (Kip, 1969). Therefore, the maximum potential modulation along the channel is obtained when one side of the device is held at a constant voltage and the other is swept to reduce the device conductance. Even so, the potential modulation induced by the biased Schottky gate along the 1D channel at the depth of the 2DEG is much smaller than that at the surface (Landauer, 1957, 1970; Kotthaus *et al.*, 1982).

The conductance of the devices as a function of gate voltage are shown in Fig. 1 when both gate fingers are swept together. Plateaux conductances close to their predicted values suggest that there is little potential drop in the bottleneck regions of either device and

that the series resistance is low. The devices resemble split gates in series and the clear quantised ballistic conductance plateaux confirms the ballistic nature of the electron transport and the non-addition of quantised ballistic resistance (Wharam *et al.*, 1992).

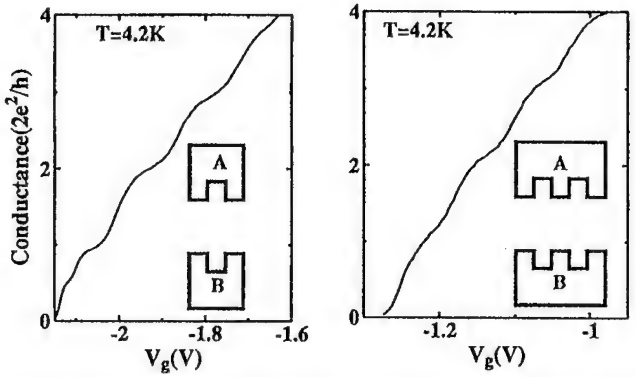


Figure 1. Conductance as a function of gate voltage V_g at 4.2 K when the gates are swept together: (a) single quantum box, (b) double quantum box. The insets show the device geometries.

In the experiment, the Fermi energy E_F is constant and an increasing negative gate voltage raises the barrier height towards E_F. In the simulation, a constant barrier shape is assumed and the incident energy E is varied. We recognise that there is a change in barrier shape as the gate voltage is swept and that we can only expect qualitative agreement with the experiment. The change in barrier shape leads to thicker barriers at lower conductance, and may account for the decrease in oscillation amplitude near pinch-off compared with the model.

The conductance was calculated for the two devices using potential profiles with 10 meV barrier height and linearly graded potential either side of the device. The single quantum box was modelled by a double barrier with the same pitch as the device (200 nm)and 1.5 meV modulation and the double quantum box was modelled by a rectangular triple barrier with 200 nm period and 1 meV modulation. The value of the potential modulation was obtained by fitting the experimental data. The well width of the double quantum box was reduced from the lithographic dimension to approximate the depletion region around the gate fingers. The conductance of the single quantum box is for the particular case when $V_g(A)$ is held at -2.6 V and $V_g(B)$ swept, and that of the double quantum box is when $V_g(A)$ is held at -1.2 V and $V_g(B)$ swept. Figures 2 and 3 show the calculated (dashed line) and experimental (solid line) conductance and the insets show the potential profile in each case.

Comparison of the single quantum box conductance with the calculated curve suggests that the lowest conductance peak is due to resonant tunnelling through a bound state in the well and the higher conductance peak is due to resonant transmission through a quasi-bound state above the well.

Examination of data in the literature suggests that the high frequency oscillations in the double quantum box conductance at low energy are due to length resonance over the entire device length (Szafer et al., 1989) and the two large dips in conductance are the beginnings of mini-gap formation associated with the periodicity of the device (van Wees *et al.*, 1991). This is in contrast to earlier work with less confinement where structure only became apparent when a magnetic field was applied device (van Wees *et al.*, 1991).

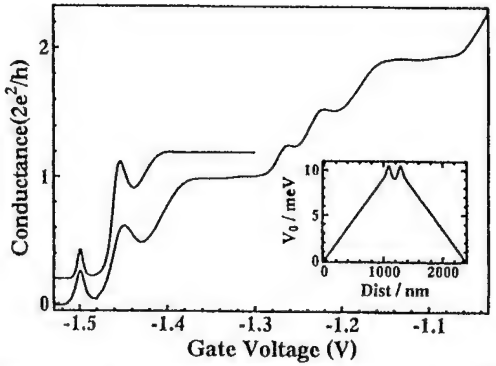


Figure 2. The offset calculated (upper) and experimental (lower) conductance of the single quantum box for the particular case when $V_g(A)$ is held at -2.6 V and $V_g(B)$ swept. The inset shows the potential profile.

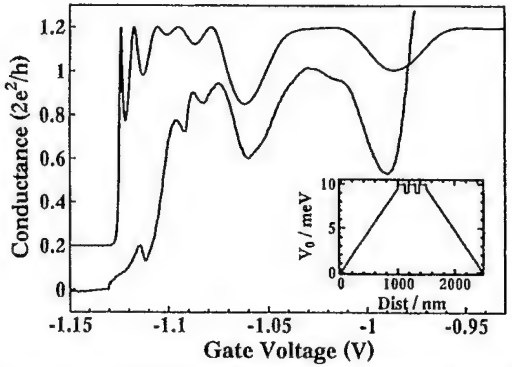


Figure 3. The calculated (dashed line) and experimental (solid line) conductance of the double quantum box for the particular case when $V_g(A)$ is held at -1.2 V and $V_g(B)$ swept. The inset shows the potential profile.

In summary, we have presented low temperature zero magnetic field conductance measurements of a single and double quantum box which are similar to theoretical predictions of phase coherent transport along 1D channels with a modulated potential. Agreement between experiment and a simple 1D barrier model is closest when there is a single occupied one-dimensional subband. Features observed on the ballistic quantised conductance plateaux of the single quantum box diminish in strength with increasing 1D subband index. Electron phase coherence along the devices is indicated at the respective measurement temperatures of 300 mK and 35 mK.

In the regime of electron occupancy in a small dot, the many body aspects of the problem must be considered. To a remarkably precise degree these can be represented by the Coulomb blockade arising from an electron charging energy e^2/C where C is a classical capacitance.

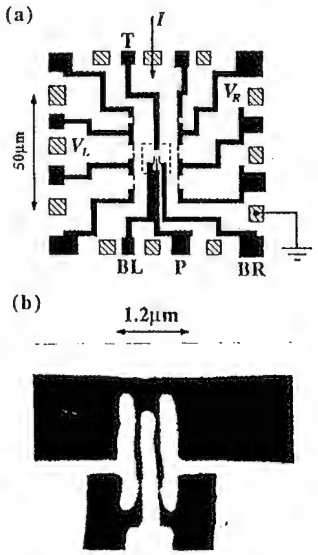


Figure 4. (a) Gate design for thermoelectric measurements on the device referred to in the text as dot *B*. Here the dark regions represent Schottky gates, the unshaded areas are regions of 2DEG, and the hatched areas represent ohmic contacts to the 2DEG. The voltage probes along the heating channel allow four-terminal resistance measurements to be made. (b) Scanning electron micrograph of the central region, as marked in (a) by a dashed rectangle.

COULOMB BLOCKADE AND SINGLE ELECTRON THERMOPOWER

As discussed previously, a patterned gate structure can be used to produce an isolated electron island (or quantum dot) separating two reservoirs of 2DEG (Smith *et al.*, 1988), and if the capacitance C of this dot is sufficiently small, single electron tunnelling (SET) effects may be observed (for a review of SET activity, see Beenakker and Staring, 1992). The two most widely studied phenomena are the Coulomb blockade (CB) oscillations in the electrical conductance G through the dot, as a function of gate voltage, and the Coulomb staircase in the dot's I-V characteristic. A remote sensor can been used to non-intrusively measure the electrical potential on the dot (Field *et al.*, 1993), providing a new method to study these effects. Understanding of the transport mechanisms can also be considerably enhanced by studies of the thermopower S. When a temperature difference ÆT is applied across an electron system, a voltage difference ÆV is set up to counteract the flow of electrons, and the ratio between these two quantities is defined as the thermopower, i.e. $\Delta V = S\Delta T$. Recent measurements (Dzurak *et al.*, 1993; Staring *et al.*, 1993) demonstrate Coulomb blockade oscillations in the thermopower of a quantum dot with the same period in gate voltage as those in the conductance. The thermopower is negative when transport through the dot is predominantly by electrons, and positive when hole transport dominates, so that S oscillates about zero as the gate voltage defining the dot is swept and if conduction is by excitation. On the other hand, if conduction is due to electrons at the Fermi energy, oscillations in resistance will be reproduced in the thermopower.

We discuss here two sets of thermopower measurements obtained on quantum dot structures with slightly differing designs, one of which is shown in Fig. 4. Schottky gates were used to define two channels of 2DEG, each 10 μm wide, on either side of a quantum dot defined using the gates T, BL, P and BR. Thermoelectric measurements were made by passing a current I along the right-hand channel to increase the electron temperature T_e in this region above the lattice temperature T_L (Long *et al.*, 1983; Syme *et al.*, 1989). The left-hand channel can be assumed to act as an electron reservoir at T_L. The temperature difference, $\Delta T = (T_e - T_L)$, between the right and left channels produces an electrochemical potential difference, $\Delta\mu_{dot} = -eS_{dot}\Delta T$, across the quantum dot of thermopower S_{dot}. Voltage probes were defined in the walls of the two channels using split-gate arrangements, and the transverse voltage $V_{tr} = (V_R - V_L)$ measured between these two probes is equal to $-\Delta\mu_{dot}/e$ plus a small, constant offset due to the thermopower of the point-contact voltage probes themselves (Molenkamp *et al.*, 1990). The measurements were performed in either a pumped 3He cryostat, or a dilution refrigerator, the latter with a base temperature of 50 mK.

Figure 5 displays typical results for the electrical conductance G_{dot} and transverse voltage V_{tr} of the quantum dot in the CB regime for one of the samples (dot *A*), as a function of the voltage V_{PC3} on one of the gates (PC3) defining the dot. A schematic of this device is displayed as an inset to the figure. These measurements were performed using a d.c. current I_{dc} in the heating channel (see inset) and the transverse voltage V_{tr} was measured using a d.c. nanovoltmeter with an input impedance of above 100 MΩ. For $V_{PC3} < -2.25$ V transmission was by tunnelling ($G_{dot} < 2e^2/h$). In this regime, oscillations in G_{dot} were observed as a function of both V_{PC2} and V_{PC3} (Fig. 5a). These oscillations were superimposed upon a rapidly rising background. From additional dc bias experiments (Dzurak, 1994) we were able to determine a value of $e^2/C \approx 2.5$ meV for the charging energy of the dot. The transverse voltage V_{tr} is plotted as a function of V_{PC3} in Fig. 4b. For $I_{dc} = -0.7$ μA, V_{tr} shows large oscillations in V_{PC3} with the same period as the CB oscillations in G_{dot}. The amplitude of these oscillations was found to increase with $|I_{dc}|$, as expected for a thermoelectric voltage, which should be proportional to ΔT. When no deliberate temperature difference was applied ($I_{dc} = 0$), a background gate-voltage dependence was still observed (chained line in Fig. 5b), which we attribute to unintentional noise heating in the sample, but this background was generally much smaller than the measured signal. We note that if T_e is very high electrons will be thermally excited over the electrostatic barriers defining the quantum dot. Activated transport by electrons always results in a negative thermopower (Cutler et al., 1969). A high proportion of activation also destroys the minima in G_{dot}, as we see in Fig. 5a for $I_{dc} = 2$ μA.

We now turn to measurements obtained more recently on a device (dot *B*) with a gate pattern as depicted in Fig. 4. In this device the two gates *BL* and *BR* were used to define the tunnel barriers at the entrance and exit of the dot, while the "plunger" gate *P* was used to raise the bottom of the conductance-band in the dot, thus removing electrons one at a time without significantly affecting the tunnel rates through the barriers. The improved regularity of the resulting Coulomb blockade oscillations is evident in Fig. 6a. The

lithographic dimensions of the dot were 0.4 μm × 0.5 μm which gives an area of about 0.12 μm^2, assuming an average depletion width of about 0.1 μm. These dimensions were confirmed from observations of Aharonov-Bohm oscillations in the conductance through the dot as a function of magnetic field in the quantum Hall regime (see, for example, Brown *et al.*, 1989). The periodicity of these oscillations indicated an area of 0.09 μm^2. The charging energy of the dot was e^2/C≈1.0 meV. This was determined from dc-bias measurements in zero magnetic field (Johnson *et al.*, 1992; Foxman *et al.*, 1993).

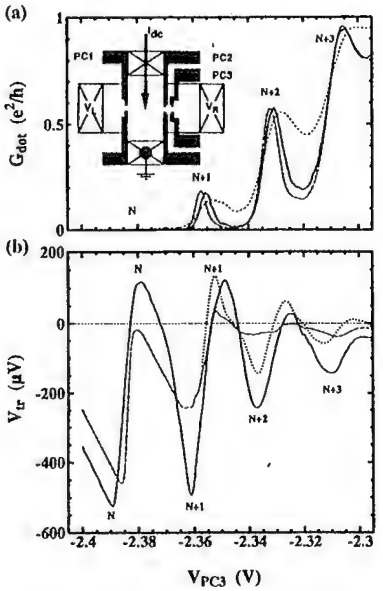


Figure 5. Results for dot *A*. (a) G_{dot} as a function of V_{PC3} with T_L=550 mK, V_{PC1}=−2.25 V, V_{PC2} =−1.65 V, and at three different heating currents: I_{dc}=0 (chained line), −0.7 μA (solid), and 2.0 μA (dashed). The peaks are marked with the inferred number of electrons in the dot: *N*, *N*+1, etc. Inset: Schematic of the device, showing the three split-gates (black) used to define the heating channel and quantum dot. (b)V_{tr} measured as a function of V_{PC3} for the conditions described above, with I_{dc}=0 (chained line) and −0.7 μA (solid); and the theoretical prediction (dotted) for V_{tr} at I_{dc}=0.7 μA, calculated as outlined in the text.

The transverse voltage across the dot was found to be a few μV when measured in d.c., and as with the previous measurements a gate-voltage-dependent background was observed even with no intentional heating in the channel. In order to extract the thermoelectric component of this signal due to the intentional heating, an a.c. current *I* was used to heat the electron gas at a frequency *f*≈5 Hz while the thermoelectric voltage V_{tr} was measured with a lockin amplifier with a reference signal set to double the frequency (2*f*). This technique has previously been used to measure the thermopower of a quantum point contact (Molenkamp *et al.*, 1992). The thermoelectric voltage occurs at twice the frequency of the heating current *I* since the temperature increase Δ*T* depends on the amplitude |*I*|, and at low powers behaves as $\Delta T \sim I^2 \sim \sin^2(t) = [1-\cos(2t)]/2$.

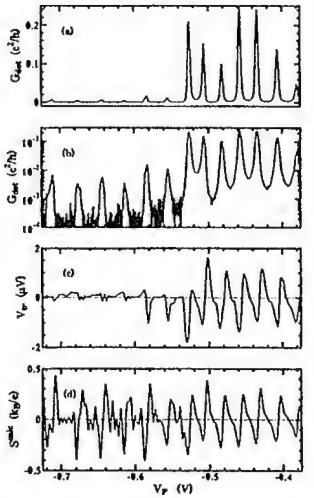


Figure 6. G_{dot} for dot *B* plotted as a function of plunger gate voltage V_P on (a) a linear, and (b) a logarithmic scale. (c) Measured transverse voltage V_{tr} across dot *B* as a function of V_P for a heating current of *I*=30 nA. (d) Theoretical thermopower S_{calc} for the case of metallic transport, calculated from (1) using the G_{dot} data shown in (a), and the parameters discussed in the text.

V_{tr} is plotted, along with *G*, log(*G*) and *d*[log(*G*)]/*dV*$_P$, as a function of the plunger gate voltage V_P in Fig. 6. For V_P>−0.52 V, there is enough broadening of the peaks to ensure that the minima are relatively temperature insensitive and do not tend to zero as $T \to 0$. This is more noticeable on the logarithmic scale. At more negative gate voltages the conductance of the dot never rises above $0.01e^2/h$ and the minima fall below the noise threshold, corresponding to a resistance in excess of 100 MΩ. When lifetime broadening in excess of the thermal energy is present electrons can tunnel directly through the dot at energies close to E_F and transport can be considered metallic. One can then think of an energy-dependent conductance *G*(*E*) which varies slowly over an energy scale of order k_BT and the thermopower is then given by (Mott and Jones, 1936)

$$S = -\frac{\pi^2}{3} k_B T \frac{k_B}{e} \frac{d[\ln G(E)]}{dE}\bigg|_{E_F}. \qquad (3)$$

The gate voltage V_P can be converted to an energy scale using $e^2/C = \alpha e \Delta V_P$, where ΔV_P is the gate voltage period of the conductance oscillations. The factor α=0.05 was obtained from separate d.c. bias experiments on the dot (Dzurak *et al.*, 1994b), following the method of Frost *et al.* (1994). The base electron temperature of *T*≈150 mK was determined by measuring the lattice temperature dependence of the sharpest conductance peaks. Note that the electron gas never reaches the base lattice temperature of T_L=50 mK due to the presence of noise heating. Using the measured values for G_{dot}, and the parameters determined above we have calculated the thermopower Scale from (1). Referring to Fig. 6, we see that in the

metallic regime when V_P>–0.52 V, V_{tr} shows good qualitative agreement with theory. The amplitude of the thermovoltage, given by V_{tr} =$S\Delta T$, then indicates that ΔT≈50 mK. This value of ΔT is consistent with values obtained from a study of the dependence of the amplitude of Shubnikov-de Haas oscillations in the heating channel as a function of both temperature and heating current, as has been performed by other workers (Ma *et al.*, 1991). Equation (1) was also found to fit the data obtained on dot *A*. The calculated result is the dotted line in Fig. 5b, which shows good agreement with the experimental data. Here ΔT=1.2 K, which was obtained from additional measurements of one of the 1D ballistic constrictions (PC2), and fitting the results to predictions for a saddle-shaped potential (Dzurak *et al.*, 1993b).

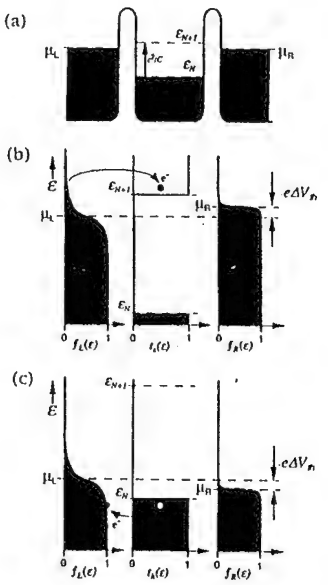


Figure 7. (a) Energy diagram for an electron island separating two electron reservoirs, with a charging energy on the island of e^2/C. Here the energy ε_N represents the maximum electron energy in the *N*-electron ground state of the system. (b) When ε_{N+1} is the closest Coulomb level to the mean Fermi energy of the system, electron transport can only occur at $\varepsilon>\varepsilon_{N+1}$, where $t(\varepsilon)$=1. Because $T_L>T_R$, there are more electrons able to tunnel from left-to-right than in the reverse direction. To maintain current balance, μ_R must rise, resulting in a negative thermopower. (c) When ε_N is the closest level, electrons can tunnel from the dot into available states in the reservoirs. Because of its higher temperature, the left-hand reservoir has more holes below μ than that on the right-hand side. To maintain current balance, μ_R must decrease, this time giving a positive thermopower.

We now discuss our results in the activated regime, where only electrons in the exponential tails of the Fermi distributions have sufficient energy to surmount the Coulomb gap. The thermopower was first calculated for this situation by Beenakker and Staring (1992) who predicted a sawtooth lineshape for the thermopower. This theory was subsequently shown (Dzurak *et al.*, 1993a) to be consistent with treating transport as activated to a Coulomb level well above the Fermi energy. A simple derivation is as follows. Consider a quantum dot which separates a left-hand reservoir of temperature (T+ΔT) and electrochemical potential μ_L, from a right-hand reservoir of temperature T and electrochemical potential μ_R=μ_L-$e\Delta V_{th}$, where ΔV_{th} is the thermoelectric voltage created by the temperature difference, ΔT. In the regime of Coulomb charging we define ε_N to be the highest occupied energy-level of a dot containing N electrons. The addition of an extra electron to the dot raises this energy by e^2/C, to ε_{N+1}=ε_N+e^2/C. For the situation depicted in Fig. 7, ε_{N+1} is the closest Coulomb level to μ_L, but the ground state of the system is the *N*-electron configuration. Electrons can only tunnel into the dot if they are thermally activated above the energy ε_{N+1}, so that the effective transmission function $t_e(\varepsilon)$ for electrons through the dot is a step-function in energy, as depicted in Fig. 6b. When ε_{N+1}–μ_L>>k_BT the electrical conductance falls exponentially to zero, however, there is still a small Boltzmann tail of electrons at energy ε_{N+1} which are capable of tunnelling into the dot. The thermopower is measured when zero current flows, so we must equate the number of electrons at ε_{N+1} in the two reservoirs, given by $f_L(\varepsilon_{N+1})$=exp[-(ε_{N+1}–μ_L)/k_B(T+ΔT)] and $f_R(\varepsilon_{N+1})$= exp[-(ε_{N+1}–μ_L-$e\Delta V$)/k_BT]. This gives a value for the thermopower, S=-$\Delta V/\Delta T$= -(k_B/e)[(ε_{N+1}–μ_L)/k_BT], which is a standard result for activated electron transport (Cutler *et al.*, 1969). Note that S<0.

When ε_N is the closest Coulomb level to the Fermi energy (see Fig. 7c), transport is due to thermally activated holes tunnelling into the dot, below the energy ε_N. The above argument again applies, except that now S=(k_B/e)[(μ_L-ε_N/k_BT], and the thermopower is positive. As the gate voltage defining the dot is swept, the thermopower reaches an absolute minimum of -$e/2CT$ when $E_c(N)$ is $e^2/2C$ above E_F, and then abruptly changes to a maximum of +$e/2CT$ as the nearest available transport level changes to $E_c(N$-1), which is $e^2/2C$ below E_F, resulting in a "saw-tooth" lineshape as a function of E_F (or gate voltage).

The amplitude of the thermopower oscillations is given by the ratio of the charging and thermal energies, ($e^2/2C$)/k_BT, multiplied by the natural unit of thermopower k_B/e≈86 μV/K. In general this is always much larger than in the metallic regime, where usually S<k_B/e. Referring to Fig. 7c we see that the metallic thermopower S_{calc} calculated from (1) is always below k_B/e. Comparing this result with the measured V_{tr} we find good qualitative agreement in the metallic regime, but in the activated regime for, V_P<–0.52 V, V_{tr} becomes smaller and does not show the sawtooth lineshape predicted by theory. It is not clear yet whether the disagreement is experimental in origin or is due to the consequences of excitation being too simple a description of an essentially many body effect.

Finally we report evidence for the first measurement of 0D excited states in the thermopower of a quantum dot. Fig. 8 shows V_{tr} at three values of heating current I, for a configuration of dot *B* with slightly different gate voltages than in the previous results. The data shows reproducible structure with a period in gate voltage about an order of magnitude smaller than the Coulomb blockade oscillations, which we believe may be attributable to transport through the discrete single-particle energy levels in the vicinity of the nearest available Coulomb level. Such structure in the thermopower has recently been predicted by theory (Beenakker *et al.*, 1992). In the raw data there was some drift of the structure in gate voltage, by up to a few mV, and we have corrected for this drift by centring on the zero-crossing of V_{tr} in the vicinity of V_P≈-0.492 V. The random shifts in effective gate voltage of the second peak may result from random telegraph signals (Cobden *et al.*, 1991). The three curves in Fig. 8 represent three different values of heating current I, and hence three different values of ΔT across the dot. The amplitude of V_{tr} increases with ΔT, as expected, but the additional structure becomes less pronounced as ΔT rises. Presumably $k_B\Delta T$ is becoming comparable to the energy-level spacing in the dot.

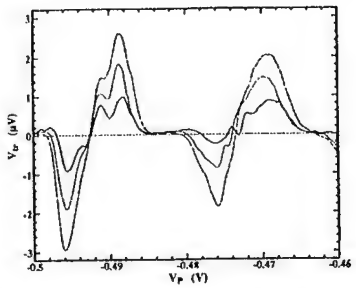


Figure 8. Measured V_B across dot B as a function of V_P, for three different heating currents in the right-hand channel: I=30 nA (solid line), 40 nA (dotted) and 50 nA (chained). As discussed in the text, the curves have been shifted slightly in gate voltage to correct for drift.

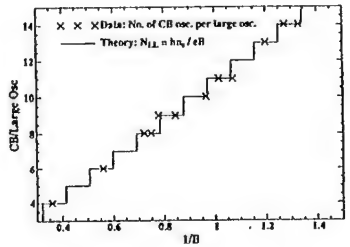


Figure 9. Data Points: Number of small CB oscillations per large oscillation period as a function of magnetic field. Line: Theoretical fit to the number of confined LL's in the dot as a function of magnetic field using an estimate of the sheet carrier density inside the dot, $n_s = 2.6\times10^{15}$ m^{-2}. Insert: A schematic diagram of the gate geometry.

COULOMB BLOCKADE AND SINGLE ELECTRON MAGNETOTRANSPORT

Applying a magnetic field to a quantum dot produces magnetic quantisation of the electron energy levels, which form Landau levels (LL). Charge transport is now by means of edge states (Fertig *et al.*, 1987) and a split gate constriction reflects consecutive edge states as it is pinched off. The magnetic field strength and the electron sheet carrier concentration fix the number of LL's present, and it is possible to tune a quantum dot to confine any number of these LL's whilst allowing charge transport via the edge states of the unconfined LL's. The edge state of the outer confined LL's may be close enough to the edge states in the leads to allow electrons to tunnel.

Using a similar quantum dot to that used in the thermopower experiments described above, a magnetic field is applied so that typically eight edge states are present in the dot. The evolution of CB oscillations is followed as the tunnel barriers are lowered allowing greater coupling to the dot.

In zero magnetic field, with both constrictions $G << e^2/h$, CB oscillations of period $\Delta V^{CB} = 12.3 \pm 0.5$ mV are observed when the plunger gate voltage is swept, allowing the capacitance of the plunger gate to be calculated. Sweeping each gate in turn allows the capacitance to be measured directly. Adding up all the capacitances, including regions of 2DEG separated from the dot by gates, gives the total capacitance $C = 2.9 \pm 0.2 \times 10^{-16}$ F and the change in electrostatic potential on adding an extra electron is $e/C = 550 \pm 30$ μV. If the barriers are relaxed ($G \sim e^2/h$) periodic oscillations in the magnetoconductance are seen with period $\Delta B = 28 \pm 1$ mT. Interpreting these as Aharonov-Bohm (AB) oscillations (Brown *et al.*, 1989; van Wees *et al.*, 1989) gives an active area of $1.5 \pm 0.1\times10^{-13}$ m^{-2}. The Shubnikov-de Haas oscillations indicate that spin splitting of LL's occurs at fields ≥ 1.1 T.

With an applied magnetic field and the conductance of both constrictions $G < e^2/h$, the CB oscillations are modulated by a longer oscillation (Fig. 10a) as observed by Staring *et al.* (1992) and Alphenaar *et al.* ((1992). The period of the large oscillation scales with the number of spin split LL's in the dot ($N_{LL} = hn_s/eB$). On increasing the temperature, the large oscillations die out at temperatures T ≥ 400 mK, whereas the small CB oscillations continue beyond T = 1 K. Figure 9 shows the number of CB oscillations per large oscillation period as a function of magnetic field for conductance $G < e^2/h$, together with a theoretical fit of the number of confined LL's in the dot.

The CB oscillations correspond to removing one electron from the dot per CB period, i.e. $\Delta N/\Delta V^{CB}$ = the reciprocal of the CB period. If the large period oscillation is attributed to AB oscillations in the outer edge state, the rate of change of area enclosed by this edge state with gate voltage, $\Delta V_t/\Delta A$, can be obtained. Combining this with the reciprocal CB period gives the local sheet carrier concentration in the dot $\Delta N/\Delta A = 2.6 \pm 0.2\times10^{15}$ m^{-2}. This value is lower than the bulk 2DEG obtained from Shubnikov-de Haas, but a remarkably good fit is found if it is used to determine the number of Landau levels present in the dot at various fields (the theoretical fit in Fig. 9).

Relaxing the barrier heights of the split-gate constrictions increases the conductance through the dot. The behaviour of the dot may be followed from just CB at very low conductances ($G<<e^2/h$), to modulated CB oscillations ($G<e^2/h$). These continue when the conductance rises above e^2/h and also $2e^2/h$, corresponding to one and two transmitted spin split edge states respectively (Taylor *et al.*, 1989).

The oscillations are modified each time another edge state is included into the charge transport. Figs. 10a and 10b show data in which the dot conductance increases from $G<e^2/h$ to $G>e^2/h$, and Fig. 10c shows data taken at $G>2e^2/h$, together with the respective power spectra.

As the conductance G increases through e^2/h (Figs. 10a and 10b), the CB oscillation evolves into two distinct periods of 13.2 and 15.2 ± 0.5 mV, with a possible small contribution of a third period at 11.7 ± 0.5 mV. There are 8 LL's in the dot at a field of 1.285 T. Above $G=e^2/h$, only $(N_{LL}-1)/N_{LL}$ of the original number of electrons are confined. The capacitance between the gate and the confined electrons will have changed by the same ratio $(N_{LL} - 1)/N_{LL}$, and the CB period should therefore change by the inverse of this ratio. Similarly if electrons are also tunnelling directly into the 3rd LL (i.e. the next confined LL) in parallel with the other transport, then a separate dot with $N_{LL} - 2$ confined LL's should be observed. The measured periods of $\Delta''V^{CB}$ are in the ratios 15.2:13.2:11.7 Å (1/6):(1/7):(1/8). The CB oscillations are interpreted as direct tunnelling into the 3rd, 2nd and 1st LLs. Note that even though the total dot conductance just exceeds $G=e^2/h$, the edge state of the 1st LL is not yet fully transmitted. The period of the 1st LL has decreased (11.7 mV above $G=e^2/h$ compared with 12.3 mV below) since relaxing the tunnel barriers allows the edge state to get a little larger increasing its capacitance to the gate.

Similarly above $G=e^2/h$. (Fig. 10c) the Fourier analysis of the data reveals three CB peaks at 14.9, 18.1 and 22.4 mV. The 14.9 mV peak corresponds to 6 confined LL's in the dot, this was at 15.2 mV when $G<2e^2/h$, a shift to a shorter period is expected as described

above. The 18.1 and 22.4 mV periods correspond to direct tunnelling into the 5th and 4th LL's respectively, with the ratios 22.4:18.1:14.9 Å (1/4):(1/5):(1/6). The 6th LL is sufficiently well coupled to the leads that CB in that LL is much reduced, possibly by one of the barriers being lower than the other.

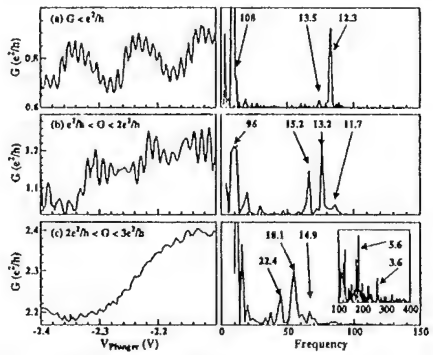


Figure 10. CB and large period oscillations for different total dot conductances with power spectra. The period in millivolts for each peak is indicated on each power spectrum. (a $G<e^2/h$. (b) $G>e^2/h$. (c) $G>2e^2/h$. Insert: the high frequency oscillations for the 4th and 5th LL (see text).

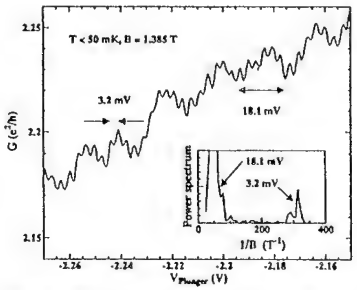


Figure 11. High frequency oscillations (of period ~ 3.2 mV) coexisting with CB oscillations of period 18.1 mV. $G>2e^2/h$. Insert: the power spectrum of the data.

In addition to the CB oscillations and the long period modulation there is another very high frequency oscillation present in some of the data at $G>2e^2/h$. These oscillations have a period of ~ 3 - 6 mV, Fig. 11 shows a situation where these oscillations co-exist with CB oscillations.

The tunnelling probability is affected by the overall charge on the inner LL's, giving rise to the long period modulation of the CB. It may also be affected by the arrangement of electrons in the inner LL's. For a CB oscillation to occur, an electron in an inner LL must jump through successive LL's until it reaches a LL from which it may tunnel into the leads. Although the charge on the dot has not changed, the electrostatic potential of a LL will vary slightly from its neighbours due to the capacitative coupling between the edge states (see Fig. 12). The tunnelling rate into and out of the outer confined LL should be sensitive to the arrangement of charges between the inner LL's, the outer confined LL is acting as a detector of the local potential generated by the inner LL's. This internal charging of edge states has also been observed in the tunnel rates of electrons from one edge state to another (Thornton *et al.*, 1986).

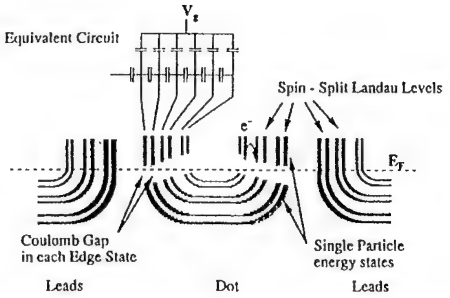


Figure 12. A schematic diagram of the energy levels of the dot and leads, including an equivalent circuit showing all the various capacitative coupling. Each edge state is made up of many single particle states and has a Coulomb gap. Electrons have sufficient thermal energy to tunnel from one edge state to another, changing the electrostatic potential of one edge state with respect to the others.

The data shown in Fig. 11 was taken at a slightly different magnetic field (1.385 T compared with 1.285 T for the previous data). There are still eight LL's present and since the conductance $2e^2/h < G < 3e^2/h$, two of the eight edge states are fully transmitted. The entrance and exit tunnel barriers were set slightly differently, and the Fourier spectrum of this data (insert to Fig. 11) shows that the 18.1 mV period CB is dominant, due to direct tunnelling into the 5th LL. If the high frequency oscillation is due to the tunnelling rate being sensitive to the arrangement of electrons on the four LL's within the edge state of the 5th LL, then the frequency should reflect the possible arrangements of charge between the 5th and all the inner LL's, i.e. five possible configurations of charge. Alternatively if tunnelling to the 6th LL is occurring, it would pick up six different arrangements of charge. The measured ratio is $18.1/3.2 \approx 5.7$.

For the data shown in Fig. 10c, the high frequency spectrum should be more complicated since there are three periods of CB oscillation present of 22.4, 18.1 and 14.9 mV, corresponding to direct tunnelling into the 4th, 5th and 6th LL respectively. Using the model suggested above, the high frequency oscillations should thus appear at 22.4/4 = 5.6 mV, 18.1/5 = 3.6 mV and 14.9/6 = 2.5 mV. The high frequency spectrum, Fig. 10c, of the data shows two distinctive peaks at the frequencies predicted for the 4th and 5th LL's. The peak for the 6th LL, if present, is well within the noise. The Fourier spectra of the data taken below $G = 2e^2/h$ does not show any clear high frequency oscillations.

At this particular magnetic field the LL's are only just spin split, i.e. the energies of the (1st, 2nd), (3rd, 4th), (5th, 6th), and (7th, 8th) LL's are almost equal and these edge states are very close to each other spatially (see Fig. 12). Therefore the coupling required to produce the high frequency oscillations will be much stronger above $G = 2e^2/h$ when tunnelling directly into the 6th LL is picking up the charge rearrangement in the 5th LL. Below $G = 2e^2/h$, the edge states involved in direct tunnelling to the leads may well be too far away for the potential variation on inner LL's to be picked up.

SUMMARY

The results discussed here and in the quoted references illustrate that the use of high mobility GaAs heterostructures grown by MBE offers a model system for the investigation of many aspects of mesoscopic physics. Use of split and patterned gates offers a means of shaping an electron gas into a desired pattern as first demonstrated in 1D and 0D by Thornton *et al.* (1986) and Smith *et al.* (1988), respectively. Although the technology is now mature, continual developments in techniques of growth will result in a flow of new phenomena for investigation.

ACKNOWLEDGEMENTS

This work was supported by SERC. D. A. Ritchie and I. M. Castleton acknowledge the support of Toshiba Cambridge Research Centre.

REFERENCES

Alphenaar, B. W., *et al.*, 1992, *Phys. Rev. B* 46:7236.
Beenakker, C. W. J., *et al.*, 1992, *Phys. Rev. B* 46:9667.
Brown, R. J., *et al.*, 1989a, *J. Phys. Cond. Matter* 1:6291.
Brown, R. J., *et al.*, 1989b, *Sol.-State Electron.* 32:1179.
Büttiker, M., 1990, *Phys. Rev. B* 41:7906.
Cobden, D. H., *et al.*, 1991, *Phys. Rev. B* 44:1938.
Cutler, M., *et al.*, 1969, *Phys. Rev. 181*:1336.
Dzurak, A. S., *et al.*, 1993a, *Sol. State Commun.* 87:1145.
Dzurak, A. S., *et al.*, 1993b, *J. Phys. Cond. Matter* 5:8055.
Dzurak, A. S., *et al.*, 1994a, submitted for publication.
Dzurak, A. S., *et al.*, 1994b, unpublished.
Fertig, H. A., *et al.*, 1987, *Phys. Rev. B* 36:7969.
Field, M., *et al.*, 1993, *Phys. Rev. Lett.* 70:1311.
Foxman, E. B., 1993, *Phys. Rev. B* 47:10020.
French, A. P., and Taylor, E. F., 1978, "An Introduction to Quantum Physics," Van Nostrand Reinhold, Workingham, UK.
Frost, J. E. F., *et al.*, 1993, *J. Phys. Cond. Matter* 5:L559.
Frost, J. E. F., et al., 1994, *Phys. Rev. B*, to be published.
Ji, Z.-L., 1993, *Semicond. Sci. Technol.* 8:1561.
Johnson, A. T., *et al.*, 1992, *Phys. Rev. Lett.* 69:1592.
Kip, A. F., 1969, "Fundamentals of Electricity and Magnetism," 2nd Ed., McGraw-Hill Kogakusha Ltd., Tokyo, 31.
Kirczenow, G., 1988, *Sol. State Commun.* 68:715.
Kotthaus, J. P., *et al.*, 1982, *Surf. Sci.* 113:481.
Kouwenhoven, L. P., *et al.*, 1990, *Phys. Rev. Lett.* 65:361.
Kumar, A., 1992, *Surf. Sci.* 263:335.
Kumar, A., *et al.*, 1989, *Appl. Phys. Lett.* 54:1270.
Landauer, R., 1957, *IBM J. Res. Develop.* 1:223.
Landauer, R., 1970, *Phil. Mag.* 21:683.
Levinson, Y. B., *et al.*, 1992, *Phys. Rev. B* 45:11936.
Long, A. P., *et al.*, 1983, *Physica B* 117:75.
Ma, Y., *et al.*, 1991, *Phys. Rev. B* 43:9033.
Molenkamp, L. W., *et al.*, 1990, *Phys. Rev. Lett.* 65:1052.
Molenkamp, L. W., *et al.*, 1992, *Phys. Rev. Lett.* 68:3765.
Mott, N. F., and Jones, H., 1936, "The Theory and Properties of Metals and Alloys," Clarendon, Oxford.
Patel, N. K., *et al.*, 1991, *Phys. Rev. B* 44:549.
Smith, C. G., *et al.*, 1988, *J. Phys. C* 21:L893.
Smith, C. G., *et al.*, 1988b, *J. Phys. C* 25:L893.
Smith, C. G., *et al.*, 1989, *J. Phys. Cond. Matter* 1:9035.
Staring, A. A., *et al.*, 1992, *Phys. Rev. B* 46:12869.
Staring, A. A., *et al.*, 1993, *Europhys. Lett.* 22:57.
Syme, R. T., *et al.*, 1989, *J. Phys. C* 1:3375.
Szafer, A., *et al.*, 1989, *Phys. Rev. Lett.* 62:300.
Taylor, R. P., *et al.*, 1989, *Phys. Rev. Lett.* 69:1992.
Thornton, T. J., *et al.*, 1986, *Phys. Rev. Lett.* 56:1198.
van Houten, H., Beenakker, C. W. J., and Staring, A. A. M., 1992, *in* "Single Charge Tunneling," Ed. by H. Grabert and M. H. Devoret, Plenum, New York, ASI 294:167.
van Wees, B. J., *et al.*, 1988, *Phys. Rev. Lett.* 60:848.
van Wees, B. J., *et al.*, 1989, *Phys. Rev. Lett.* 62:2523.
van Wees, B. J., *et al.*, 1991, *Phys. Rev. B* 43:12431.
Wharam, D. A., *et al.*, 1988a, *J. Phys. C* 21:L209.
Wharam, D. A., *et al.*, 1988b, *J. Phys. C* 21:L887.

CONTRIBUTED ABSTRACTS BY ATTENDEES

1. Nonlinear dynamical response of double-barrier resonant-tunneling structure, V.V. Afonin, A.M. Rudin, *A.F. Ioffe Physico-Technical Institute, 194021 St. Petersburg, Russia.* The nonlinear dynamical response of double-barrier resonant-tunneling structure (DBRTS) on external a.c. bias is studied. The nonequilibrium Green's function technique is exploited to derive the rate equation for the occupation number of level in the DBRTS well for the case of coherent tunneling. The conditions required for this equation are also derived. This rate equation is analyzed in details and the analytic (that is in a form which does not require numerical analysis) results for the current through the structure in all limiting cases of interest are obtained. In addition, it is shown that in the constraints of semiclassical rate equation approach there is no difference for the dynamical response problem between coherent and sequential tunneling processes, at least for such experimentally measurable quantities as level distribution function and current through DBRTS.

2. Beating pattern in the magneto-oscillations of the 2DEG at semiconductor quantum wells, E.A. de Andrada e Silva, G.C. La Rocca and F. Bassani, *Scuola Normale Superiore, Pizaaz dei Cavalieri 7, 5600 Pisa, Italy.* A beating pattern in the amplitude of the Shubnikov-de Haas oscillations of the two-dimensional electron gas (2DEG) at semiconductor quantum wells has been observed and assigned to the spin-orbit spin-splitting in the presence of inversion asymmetry. We study such beating pattern with a calculation of the electronic states and the magnetization of the 2DEG at III-V modulation doped heterojunctions as a function of the applied magnetic field and carrier concentration. Full account of the spin-orbit coupling is taken. A semiclassical analysis is used to interpret the quantum mechanical results. A regular beating pattern is shown to result from isotropic spin-splittings and an anomalous one, connected to the magnetic breakdown at the points of the Fermi "surface" with a small splitting, to result from anisotropy in the spin splitting. Specific results for heterojunctions of InAs and GaSb grown on appropriate larger gap materials, shown both above effects, are presented. The origin of the anisotropy in the spin splitting is discussed.

3. Dynamical Screening of Optical Phonons by Electrons in Quantum Wires, L.S. Bokacheva, S.V. Gantsevich, and V.L. Gurevich, *A. F. Ioffe Physico-Technical Institute, 194021 St. Petersburg, RUSSIA.* We investigate the interaction of optical phonons in the vicinity of the quantum wire with plasma oscillations in the wire. This intgeraction leads to the formation of coupled plasmon-phonon modes. The dispersion relations of these modes are found for the case of Boltzman and Fermi statistics. In the case of Fermi statistics, there are a number of plasmon modes corresponding to different transverse energy levels and therefore the spectra of plasmon-phonon modes are much more complicated than that for Boltzman statistics. We wish to emphasize that in quantum wires all kinds of plasmon spectra must have an acoustic-like (linear) dependence on the wave vector. Unlike 2D and 3D systems, where electrons can be considered as free, in the ID case one cannot neglect the Coulomb forces among electrons even in the long-wave limit. Therefore the electrons represent a liquid rather than a gas and the validity of the RPA in getting plasmon dispersion relations needs to be justified. Nevertheless it seems that the RPA calculations are in good quantative agreement with experimental data. We think that this may be understood as the fact that in a quantum wire one deals not with free electrons but with elementary electron-like excitations in a liquid with the energy spectrum renormalized by strong electron-electron interaction. We analyze the role of electron-electron interaction using the simple kinetic diagram technique for density matrix.

4. Coulomb Quantum Kinetics on the Femtosecond Time-Scale, Karim El Sayed, L.

Bányai, and H. Haug, *Institut für Theoretische Physik, Universität Frankjfurt, Robert-Mayer-Str. 8, D-60054 Frankfurt a.M., Germany.* During and shortly after a femtosecond optical laser pulse excitation of a semiconductor the carrier kinetics presents new features. At this early stage the energy broadening, due to the short time interval, is comparable to the average carrier kinetic energy and to the plasma frequency. We will show that screening is ineffective under these conditions and give a description of the build-up of screening with time. Also the kinetic equation, which describes this new quantum-kinetics, will be presented. Numerical simulations show that this initial kinetics manifest itself in a pronounced non-exponential decay of the optical polarization.

5. Ultrafast Coherent and Incoherent Dynamics in Photoexcited Semiconductors, Stefan Haas and Fausto Rossi, *Philipps-Universität Marburg, Fachbereich Physik und Zentrum für Materialwissenschaften, Renthoff 5, D-35032 Marburg, Germany,* and Tilmann Kuhn, *Institut für Theoretische Physik der Universität Stuttgart, D-70550 Stuttgart 80, Germany.* The dynamics of an optically excited semiconductor is usually described in terms of the semiconductor Bloch equations. Our solution of this coupled set of quantum kinetic equations is based on a decomposition of such equations in a coherent and an incoherent part. The former is integrated directly while the latter is solved by means of a generalized Monte Carlo simulation. The main peculiarity of the method is to retain the big advantages of the Monte Carlo method in treating scattering processes and, at the same time, to take into account on the same kinetic level also coherent phenomena. In addition to a simulation of the distribution functions for electrons and holes, this results in a simulation of the interband polarization induced by the coherent laser light field. Such an approach allows a self-consistent description of the carrier photogeneration process. The energy broadening due to the finite pulse duration and due to the decay of the interband polarization has not been introduced as a phenomenological paramater as in any conventional Monte Carlo simulation but it comes out self-consistently with its full time dependence. Furthermore, band-renormalization and excitonic effects are included in a self-consistent way. Our physical model includes both carrier-carrier and carrier-phonon interaction. Therefore, carrier thermalization due to carrier-carrier scattering and energy relaxation due to the emission of optical phonons are included on the same kinetic level. Our results show a transition from the high-density regime, where the dynamic is dominated by carrier-carrier interaction, to the low-density one, where the dynamic is dominated by carrier-phonon interaction. Especially in the low-density limit we see a strong cancellation between in and out scattering terms in the equation for the interband polarization. As an application we examined spectrally resolved band-acceptor luminescence. We obtain an excellent agreement between the measured and calculated spectra.

6. Electron Transport Phenomena in Ultra-Short Device Geometries, Julie A. Kenrow, *Dept. of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA. 94720.* As device feature sizes approach 0.1 μm, it is becoming increasingly necessary to develop time-dependent, quantum mechanical electron transport models. These models must be capable of accounting for several features, including non-stationary transport and quantum confinement along and/or across a device, quasi-ballistic transport, the intra-collisional field effect, delayed scattering due to the time-energy uncertainty relation, and both intersubband and intrasubband scattering with a non-thermal-equilibrium distribution of phonons. A model is presented here that accounts for all of the above features, and treats a single electron and the device LO and SO lattice phonons as a coupled system that is time evolved from a particular initial condition at 0 K, through direct solution of the Schrödinger equation. The time evolution of both the electron wavepacket and the phonon mode populations are presented, primarily in quantum wire geometries, for various applied field strengths. The effectiveness is intrasubband and intersubband scattering in affecting electron transport is studied in a single

wire for an electron initially placed in a specified eigenstate of the GaAs/$As_xGa_{1-x}As$ wire system. Preliminary results of resonant intersubband electron-SO-phonon scattering studies are presented both in single wires and in coupled asymmetric wire geometries.

7. Electronic Transport in Ultrathin Wires with Near Atomic Diameters, Y. Kumzerov, *Physico-Technical Institute Russian Academy of Sciences, 194021 St. Petersburg, Russia.* This work is related with fabrication and investigation of some physical properties of materials in the form of wires (Hg, Sn, In, Cd, Bi, Pb, Ga, S, Se, Te, InSb) with diameters which can be successively varied from 2 to 15 nm. These wires are thinnest among the wires created and studied at this time. Such diameters, which are unapproachable for the modern lithographic technique was obtained thanks to nontrivial approach to this problem. There were used the natural dielectric matrixes obtaining the regular set of identical, parallel, isolated channels with near atomic diameters which were filled under high pressure by the melted liquid material. In this work there were ob served the successive appearance and development of specific physical properties intrinsic in one-dimensional systems and therefore there was investigated at first time the transitional region between the one-dimensional and three-dimensional objects. There was investigated the electronic transport in such systems (weak localization and Peierls phase transition). There was observed the successive phase transition destruction (superconductivity and melting) with approach to the one-dimensional limit. There are presented unique now for one-dimensional objects the superconducting critical temperature and melting-freezing temperature dependencies versus wire diameter (up to 20A).

8. Temperature Dependence and Hot Electron Scattering in Heavily P-Doped GaAs: Hole Plasma and LO-Phonon Coupling, V. Narayan[1], J.M. Rorison[2] and J.C. Inkson[1], *[1]Dept. of Physics, University of Exeter, Exeter EX4 4QL, and [2]Sharp Laboratories of Europe Ltd, Oxford Science Park, Oxford OX4 4GA.* The mean free path of a single hot electron interacting with isotropic p-doped GaAs is calculated as a function of doping density and temperature within the Born approximation. The results show that there is significant temperature dependence for densities less than $10^{19}cm^{-3}$. The interaction between light hole plasmons, heavy hole plasmons and LO-phonons is studied by using a finite temperature approximation to the full finite temperature, frequency and wavevector dependent random phase approximation (RPA) for the dielectric function. For GaAs, where there is a large difference in effective masses between the bands, we find that the plasmon mode associated with the light hole band is effectively screened out at small q by the higher frequency plasmon mode associated with the heavy hole band. The interaction between the holes and LO-phonons is highly dependent on the doping density.

9. Finite Fields and Intervalley Scattering in a Resonant Tunnelling Quantum Wire and Dot Double Barrier Structure Using a Multiband Microscopic Layer Method, V. Narayan, G.P. Srivastava and J.C. Inkson, *Dept. of Physics, University of Exeter, Stocker Road, Exeter EX4 4QL.* We present a multiband calculation of resonant electron tunnelling through quantum wires and dots consisting of a double barrier structure. A layered version of the empirical pseudopotential method is used with the addition of hard wall boundary conditions to model a structured single layer (wires) and a structural individual wire (dots). Our results show clearly the importance of non-parabolicity and intervalley mixing in these reduced dimensional structures. The inclusion of electric fields allows for the investigation of Stark effects in wires and dots.

10. Phonon-Assisted Ballistic Resistance, V.B. Pevzner, *North Carolina State University, Raleigh, N. C.* There is a (negative) fraction ΔG of the ballistic conductance G that is due to electron-phonon interactins. For the special case of a spatially uniform conductor, ΔG depends upon temperature, cross-sectional area, and the chemical potential. With increasing temperature, phonon

and other effects eventually result in fully collision-controlled transport. ΔG is especially large at lower temperatures just when increasing chemical potential introduces one more conductance channel.

11. Quantum Chaos in Antidot Superlattices, P. Rotter, H. Silberbauer, M. Suhrke and U. Rössler, *Institut für Theoretische Physik, Universität Regensburg, D-93040 Regensburg, Germany.* We have studied the statistical properties of the energy spectrum of an antidot superlattice subject to a perpendicular magnetic field. Numerical studies of the classical dynamics of this system have shown a transition from chaotic to regular behavior with increasing strength of the magnetic field. According to the predictions of quantum chaos this transition should be reflected in the statistics of the energy spectrum. Our quantum mechanical calculations make use of the magnetotranslation group, according to which the eigenstates are characterized by a magnetic wave vector. As a function of this wave vector, in the regime of low magnetic field, we are able to detect a transition between different universality classes (Gaussian Orthogonal Ensemble - GOE, Gaussian Unitary Ensemble - GUE) corresponding to different point group symmetries in the magnetic Brillouin zone. The magnetic field dependent transition from a Wigner to a Poisson type of statistics in the distribution of nearest neighbor spacings is shown in the upper part of the figure for a fixed magnetic wave vector of low symmetry. This is in accordance with the transition from chaotic to regular motion in classical dynamics. Longer ranged fluctuations have been analyzed for the same spectra by using the spectral rigidity (Δ_3-statistics) and reveal the same transition as can be seen in the lower part of the figure.

12. Direct Patterning of Silicon Dioxide Films Using Vapor Etching of Beam-Induced Carbon, J.M. Ryan, J. Allgair, T.K. Whidden, M.N. Kozicki, and D.K. Ferry, *Center for Solid State Electronics Research, Arizona State University, Tempe, AZ 85287-6206.* We present results of a direct patterning technique of SiO_2 using electron beam deposited carbon and subsequent HF vapor etching. This method has the advantage over previous direct patterning methods (oxide charging, oxide damage or contamination resist and subsequent ion milling), in that the dose needed is as small as $1\mu C/cm^2$, as much as 10^6 times smaller than needed for other methods. Additionally, selectivity ratios (etch rates of exposed areas : unexposed areas) as high as 100 have been achieved, as compared to oxide damage etch enchancement of only three to four times. The high selectivity of CEVE is accomplished in the etch step, by utilizing the volatility difference of beam-deposited carbon and ambient carbon-containing surface contamination. This large etch enhance-ment is accomplished at doses of only 100 $\mu C/cm^2$ (roughly that needed to expose conventional electron beam resists such as PMMA). While the sensitivity of CEVE is an improvement over other direct SiO_2 techniques, the resolution has not yet matched these other methods. Using this carbon enhanced vapor etching (CEVE) technique allows the fabrication of continuous lines 100 μm long under 100 nm in width to be etched in SiO_2 films as thick as 150 nm. Using an STM to deposit contamination, etched lines under 10 nm in width have been demonstrated. Thus the current resolution of 80 — 100 nm is not the final resolution of SEM exposure CEVE, and work is currently under way to optimize the SEM exposure process to determine the ultimate resolution.

13. Three-Barrier Tuned Structure as a Phonon Spectroscopy Device, V.I. Kozub and A.M. Rudin, *A.F. Ioffe Physico-Technical Institute, 194021 St.-Petersburg, Russia.* We suggest the three-barrier quantum-size tunneling structure (TBS) for the frequency-resolved acoustic phonon detection and generation. We show that making use of acoustic-phonon-mediated *interwell tunneling* transitions in TBS do provide this possibility. The efficiency of the phonon-mediated interwell transitions is analyzed in detail for the case of the quantum-dot structure and for the case of the quantum-well one both in the absence and in the presence of the quantizing magnetic field. The relation-ships between the phonon distribution

function and the current through TBS (in the phonon detection regime) and between the current and the phonon creation rate (in the phonon generation regime) are found. It is shown that if the dominant mechanism for the charge transfer corresponds to phonon-mediated transitions, the TBS has rather good parameters as a phonon spectroscopy device. The best spectral resolution is achieved for a quantum-dot structure, while for a quantum-well structure some broadening due to intrawell electron motion exists. However, a strong quantizing magnetic field, which suppresses this motion, provides a substantial improvement of spectral resolution. In addition, this device can be easily made tunable by the electro-static gate, which changes the relative asymmetry of the wells.

14. 3D Quantum Transport Simulations, D. Z.-Y. Ting, S.K. Kirby, and T.C. McGill, *Thomas J. Watson, Sr. Laboratory of Applied Physics, California Institute of Technology, Pasadena, CA 91125.* Quantum transport in mesoscopic devices is examined with an exactly solvable real-space three-dimensional supercell model. The flexibility of our model has enabled us to include elastic scattering effects due to impurities, interface roughness, and alloy disorder in our studies of 2D (double barrier hetero-structures), 1D (quantum wires electron waveguides), and 0D (quantum dots) mesoscopic device structures. Our studies reveal that structural imperfections can not only produce additional scattering processes in a perturbative sense, under the right circumstances, they can also substantially alter the quantized electronic states, leading to modified transport properties. For example, we have demonstrated that interfacial inhomogeneities in double barrier resonant tunneling diodes can induce lateral localization of wave functions; strongly attractive impurities can produce additional transmission resonances; clustering effects in alloy barriers can reduce barrier effectiveness and surface roughness in quantum dots can cause large fluctuations in transmission characteristics. In addition, we will also discuss the application of our method to arrays of mesoscopic devices, where transport properties can be influenced by coherence among closely-spaced device structures.

15. A Transfer-Matrix Approach to Photon-Assisted Transmission Through an Oscillating Quantum Well, Mathias Wagner, *Hitachi Cambridge Laboratory, Madingley Road, Cambridge CB3 0HE, United Kingdom.* The effect of enhanced tunneling in the presence of an electric field is termed photon-assisted tunneling. Typically, the presence of photons means that in addition to the usual transmission channel at energy E, side channels will open up at energies $E+n\hbar\omega$ where n is an integer. Within the framework of the transfer-matrix approach, we report on analytical and numerical results for the transmission probability through a strongly driven double-barrier structure. Two scenarios are studied: (1) A central quantum well oscillating as $V_1 \cos \omega t$ [1], which can be experimentally realized by using fine gates, and (2) an oscillating electric field applied across the structure, which is characteristic for IR radiation. In both cases we find a strong quenching or "coherent destruction" [2] of the transmission probability in all channels at particular values of $V_1/\hbar\omega$. This effect is related to the temporal phase coherence of the wave function.

[1] M. Wagner, to be published in Phys. Rev. B (June 15, 1994).

[2] M. Holthaus, Z. Phys. B **89**, 251 (1992); M. Holthaus and D. Hone, Phys. Rev. B **47**, 6499 (1993).

16. Evaluation of the Mobility in a Si-SiO_2 Inversion Layer at T=0 K Using Green's Function Formalism[1], Dragica Vasileska-Kafedziska, Paolo Bordone[2], Terry Eldridge and David K. Ferry, *Center for Solid State Electronics Research, Arizona State University, Tempe, AZ 85287-6206,* and [2]*Dipartimento di Fisica ed Instituto Nazionale di Fisica della Materia, Universita di Modena, Via Campi 213/A, 41100 Modena, Italy.* We study transport properties of a (100) Si-inversion layer at zero temperature using Green's function formalism[1]. We give the results for the density of states function and mobility for various fitting parameters and different effective

fields. The dependence of mobility on the inversion charge density provides information for the strength of the considered dissipative mechanisms. The position of the subband minima and electron wave-functions are obtained by the self-consistent solution[2-5] of the Poisson, Schrödinger and Dyson equations for each value of the effective transverse electric field. In our model, we have included both, charged-impurity and surface-roughness scattering. Scattering associated with Coulomb centers is separated in contributions from the depletion and interface-trap charge. Since the charged centers near the plane of the 2D-electron gas contribute effectively to the total broadening of the electronic states for low inversion charge densities, this scattering mechanism is treated properly using the screened Coulomb interaction[6]. For the sake of simplicity, we assume that the electrons are scattered by randomly located but identical δ-function impurity potentials that arise from the depletion charge. Surface-roughness is included via the random potential term added to the free-electron Hamiltonian of the system. We have used both, Gaussian and exponential forms for the autocovariance function[7] of this random potential term in our analytical calculations. We give the analytical expression for the broadening of the electronic states in each subband, and the expression for the conductivity that includes the correction due to the normal particle-hole ladder diagram. In addition, we present the numerical results for the mobility for fixed fitting parameters and various oxide fields. We finally give the comparison between the mobility curves for the two models of the surface-roughness autocorrelation function. The results for the mobility are in agreement with the experimental results of Kawaji obtained at 4.2K.

[1] D. Vasileska-Kafedziska, P. Bordone, and D.K. Ferry, submitted for publication.

[2] T. Ando, A.B. Fowler, and F. Stern, Reviews of Modern Physics **54**, 437 (1982).

[3] F. Stern, Phys. Rev. B **5**, 4891 (1972).

[4] B. Vinter, Appl. Phys. Lett. **44**, 307 (1984).

[5] S.E. Koonin and D.C. Meredith, *Computational Physics* (Addison- Wesley, NY., 1990).

[6] B. Yu-Kuang Hu and S. Das Sarma, Phys. Rev. B **48**, 5469 (1993).

[7] S.M. Goodnick, D.K. Ferry, and C.W. Wilmsen, Phys. Rev. B **32**, 8171 (1985).

17. General Conditions for Stability in Bistable Electrical Devices with S- or Z-shaped Current Voltage Characteristics, Andreas Wacker and Eckehard Schöll, *Institut für Theoretische Physik, Technische Universität Berlin, Hardenbergstr. 36, 10623 Berlin, Germany.*

A typical feature of semiconductor heterostructure elements is the fact that the transport of charge is impeded perpendicular to the barriers. This leads to charge accumulation in parts of the sample, which determine the electric potential. As the electric potential distribution strongly influences the transport properties, a condition for self-consistency arises. In many cases this condition allows for more than one stable solution of the charge distribution for a given bias voltage across the sample. Thus, we find a bistability in the current-voltage characteristic. Examples are the Heterostructure Hot-Electron Diode, Double Barrier Resonant Tunneling Diodes (DBRTD), quantum-dot structures, and highly doped superlattices where even 4 stable states have been found both experimentally and theoretically for a fixed bias[1]. In this contribution we present a general description for the stability of the different branches in the current-voltage characteristic. The main idea is the fact that there is an intrinsic degree of freedom, which should be treated as an independent dynamical variable (e.g. the charge stored in a quantum well). For general circuit conditions the voltage across the sample represents an additional dynamical variable. This dynamical system can be easily treated with the methods of nonlinear dynamics. Thus, some general statements can be made regarding the stability and the dynamical behavior without any knowledge of the specific transport processes. From this we obtain quite simple conditions for the stability of the different branches for both S- and Z-type characteristics. The different types of instabilities leading to oscillations, switching or current-filament formation are discussed. We illustrate the

application of the general formalism by various semiconductor devices. For example, it has been demonstrated recently that the middle branch of the Z-shaped characteristic of the DBRTD can be stabilized by an appropriate technique[2]. The under-lying principle is much easier to under-stand within our general description.

[1] J. Kastrup, H.T. Grahn, K. Ploog, F. Prengel, A. Wacker and E. Schöll, submitted to Appl. Phys. Lett. (1994).

[2] A.D. Martin, M.L. F. Lerch, P.E. Simmonds, L. Eaves and M.L. Leadbeater, Proceedings of 8th Conf. on Hot Carriers in Semiconductors (1993).

18. Coherent Transport Through an Array of Quantum Dotsf with Strong Coulomb Correlations, C.A. Stafford and S. Das Sarma, *University of Maryland, Dept. of Physics, College Park, MD. 20742.* We investigate the linear response conductance through an array of GaAs quantum dots using an SU (N) (N = 2 - 6) Hubbard model to account for the effects of quantum confinement, interdot tunneling, and strong intradot Coulomb interactions. Interdot tunneling splits the resonant tunneling peaks into minibands which are separated by energy gaps arising due to collective Coulomb blockade[1]. We analyze the scaling with system size of these energy gaps and of the optical response of the system in order to determine the locus of the Mott-Hubbard metal-insulator transition in the one-dimensional *SU* (N) Hubbard model. We argue that the experimental conductance spectra of Kouwenhoven *et al.*[2] provide evidence for the occurrence of this "metal-insulator transition" in an array of quantum dots. In a magnetic field, certain resonant tunneling peaks are strongly suppressed due to a density-dependent quantum phase transition[1] which represents spin-polarization of the high-lying electrons in the array. There is also a sharp onset of optical absorption at this phase transition, which could lead to the possibility of a magneto-optical or electro-optical switch.

[1] C.A. Stafford and S. Das Sarma, Phys. Rev. Lett. (in press).

[2] L.P. Kouwenhoven *et al.*, Phys. Rev. Lett. **65,** 361 (1990).

19. Method to Explore the Random Potential in 2DEG by a Four-Gate Device, E.V. Sukhorukov and I.A. Larkin, *Institute of Microelectronics Russian Academy of Sciences, 142432 Chernogolovka, Moscow District, RUSSIA.* Although the attraction of a 2DEG in a heterostructure is its long mean free path, which implies weak scattering, it has become clear that the random potential from the donors plays a major role in various phenomena such as a breakdown of a conductance quantization of microjunctions, mesoscopic conductance fluctuations, Coulomb blockade phenomenon in quasi one-dimensional channels, quantum Hall effect and so on. Moreover, the significance of the random potential is that it limits the performance of devices at low temperature, where impurity scattering dominates. It has proved difficult to model these tiny devices accurately because the behavior of the donors is uncertain, particularly the degree to which ionized donors are correlated in space. A direct measure of the potential due to the impurities would fix the degree of correlation and be of great benefit to modelling. We propose a new method to explore the random potential in the vicinity of a quantum point contact (QPC). The position and parameters of the saddle point of the smooth potential in the QPC are varied by four gates above the two-dimensional electron gas. Two of them form the QPC. The other two gates are placed in the source and drain areas to drive the saddle point *along* the channel. A resonant tunnelling peak appears in the conductance when the saddle point and a local minimum of the random potential coincide. The most pronounced peaks appear in the pinch-off regime when the QPC is classically closed and the Fermi energy of the two dimensional electron gas equals the binding energy in a local minimum. In contrast with the paper [1], in the present work any assumptions concerning a model of the heterostructure has not been considered. The only simplification is that we have assumed the case of narrow channel in four-gate device. A three-dimensional electrostatic problem has been solved for the above situation. The position and other parameters of the saddle point are calculated

analytically and a practical procedure for determining the parameters of the bound states is proposed.
[1] I.A. Larkin and E.V. Sukhorukov, Phys. Rev. B **49**, 5498 (1994).

20. Hole Transport in GaAs, R. Scholz, *Classe di Scienze, Scoula Normale Superiore, Piazza dei Cavalieri 7, 56126 Pisa, ITALY.* Starting from a 8*8-kp band structure, the scattering rates between hole subbands can be calculated with realistic wavefunctions. Pronounced differences to published scattering rates are found [1], partially due to the density of states of the nonparabolic light hole band and to different overlap between the wavefunctions. The scatter-ing rates are used in a model of hot displaced Maxwellians to investigate hole transport at moderate electric fields. The agreement with drift velocity measurements is good over the field range investigated. Contributions of different scattering channels to the mobility are analyzed and compared to other calculations.
[1] M. Costato and L. Reggiani, Phys. Stat. Sol. B **58**, 471 (1973).

21. Coherent Dynamics and Terahertz-Emission from Asymmetric Double Quantum Wells, E. Binder, T. Kuhn, and G. Mahler, *Institut für Theoretische Physik, Universität Stuttgart, Pfaffenwaldring 57, D-70550, Germany.* We present a theoretical investigation of the coherent dynamics in an $Al_zGa_{1-x}As$ asymmetric double quantum well structure optically excited by a short laser pulse. In such structures, states in the different wells can be brought into resonance by an applied electric field. In this case, a short laser pulse excites a spatially oscillating carrier wave-packet which yields an oscillating dipole moment and thus, as observed experimentally [1[, leads to the emission of electromagnetic radiation in the terahertz range. The calculations are based on the envelope-function formalism in effective-mass theory for the case of two bands, the conduction band and the heavy-hole band. The dynamics has been analyzed by a full numerical solution of the semiconductor Bloch equations for a multi-subband system. Here, the relevant variables are given by the distribution functions of electron and hole subbands, the inter-band polarizations, and the intraband polarizations. The oscillating dipole moment is obtained from the intraband polarizations and its second derivative gives the electric field strength of the emitted radiation. We find that the details of the emission process sensitively depend on the interplay between interband and intraband polarizations as well as on the corresponding dephasing rates. There-fore, the results demonstrate the importance of an approach which includes bound and unbound excitonic states as well as the dynamics of the intraband polarizations.
[1] H.G. Roskos *et al.*, Phys. Rev. Lett. **68**, 2216 (1992).

22. Linearized Quantum Transport Equations in *ac*-fields, Petr Král, *Institute of Physics, Academy of Sciences, Na Slovance 2, 18040 Praha 8, Czech Republic.* A new formalism is developed for evaluation of transport coefficients in quantum many particle systems. The method is built up on a systematic linearization of nonequilibrium Kadanoff-Baym transport equations in external *ac*-fields. Easy and systematic corrections can be performed in this simple set of transport equations, which replace complicated vertex equations from a Kubo formula. The method has been tested on an *ac*-conductivity of a quantum wire with an electron-phonon interaction [1]. Nonadiabatic systems, which go beyond the Kubo formalism [2], can be directly handled.
[1] P. Král, presented on *8th International Winterschool on New Developments in Solid State Physics, "Interaction and Scattering Phenomena in Nanostructures",* (Mauterndorf, Austria, Feb. 14-18, 1994).
[2] M. Wagner, Phys. Rev. B45, 11595 and 11606 (1992).

23. Carrier Transport in Quantum Well Lasers: A Comparison Between Different Heterostructures, A. Greiner and T. Kuhn, *Institut für Theoretische Physik and Synergetik, Universität*

Stuttgart, Pfaffenwaldring 57, D-70550 Stuttgart, Germany, and H. Hillmer, S. Hansmann, and H. Burkhard, *Forschungszentrum der Deutschen Bundespost Telekom, Postfach 10 00 03, D-64267 Darmstadt, Germany.* In a semiconductor laser an electrically injected carrier density yields an optical output. For high frequency applications an efficient and fast response on the injection is important. This depends on both electronic and optical properties of the semiconductor. The photon density reacts on the stepwise increase of the injection current with a characteristic delay, the turn-on delay time t_{on}, which should be as small as possible. This delay time depends strongly on the coupling between the electronic system and the light field, in particular the overlap between the carrier states and the field mode. However, from many experiments it has become evident that for an operation in the GHz regime also time-scales related to the electron and hole dynamics play a role due to the fact that these carriers have to be transported from the contacts to the active layer. To take into account these effects, the two-level rate equation model has been extended to a now widely used three-level model [1], where all transport effects are summarized by one effective transport time. Such a phenomenological description allows one to estimate the role of transport processes from a measurement, but it is not possible to make a quantitative prediction for a specific laser structure. To overcome this drawback, the transport of the carriers as well as the coupling to the light field have to be taken into account self-consistently. The aim of this work is to investigate the influence of structural parameters on the switching behavior of a quantum well laser diode. The injection current is increased stepwise from a value I_{low} to I_{high}, both being above the threshold. Carrier transport is described by a drift diffusion model including the influence of Fermi-Dirac statistics, and Poissons equation for the electrostatic potential. For the photon number we use a rate equation. Carrier densities and photon number are coupled by stimulated emission taking place in the active quantum well region. This quantum well is embedded in a wave-guide layer made by material with an energy gap between the values of the cladding layers and quantum wells. The transition between cladding layer and wave-guide zone can be abrupt (separate confinement hetero-structure, SCH) or graded (graded index SCH, GRINSCH). We are particularly interested in the role of the grading as well as the position of the quantum well on the optical response. We find that the transient properties, e.g., amplitude and frequency of the relaxation oscillations, depend on the specific structure. Since the parameters describing the coupling between the carriers and the light field are kept constant, these changes are caused by carrier transport effects. In particular it turns out, that the response becomes faster if the quantum well is shifted towards the p-doped cladding layer, which can be understood from the strongly different mobilities of electrons and holes. Furthermore, the optical response is influenced by the built-in fields produced by a grading.

[1] R. Nagarajan, T. Fukushima, M. Ishikawa, J.E. Bowers, R.S. Geels and L.A. Coldren, IEEE Photonics Tech. Lett. 4, 121-123 (1992).

24. Acoustic Phonon Relaxation in Valence Band Quantum Wells, G. Edwards*, E.C. Valadares#, F.W. Sheard# and D.K. Ferry*, **Center for Solid State Electronics Research, Arizona State University, Tempe, Arizona 85287-6206, #Department of Physics, University of Nottingham, Nottingham NG7 2RD, U.K.* The hole-phonon (h-p) interaction in semiconductor heterostructure systems plays a fundamental role in determining transport properties and hence also device applications. In a Monte Carlo simulation of a quantum well (QW) based device, knowledge of the h-p scattering rates is an essential input. In contrast to the conduction band (cb) case where the electron-phonon (e-p) interaction can be treated simply, the h-p interaction in a QW system is very complicated [1]. This is because the hole QW subbands and the h-p interaction require a multiband treatment [1]. The QW interfaces cause strong bulk light hole/heavy hole mixing in the hole subbands and this mixing also strongly affects the h-p interaction [1]. Describing the phonon system also becomes a difficult problem if anisotropy is included [1]. We have set up a

general formalism, based on the 6×6 multiband Luttinger Hamiltonian and the corresponding Bir-Pikus deformation potential matrix, for calculating acoustic phonon emission rates in GaAs/AlGaAs valence band (vb) QWs [1]. Fermi's Golden rule is employed to calculate the scattering rate [1]. The acoustic phonons are treated as bulk like and in terms of elasticity theory, with anisotropy effects retained rather than the usual isotropic approximation. This formalism extends previous work [1] by incorporating the spin split off band and warping effects. Thus narrow QWs, where the spin split off band becomes important, can also be tackled by the present method. We have examined the acoustic phonon inter-subband relaxation rate, between the lowest two subbands, from a fixed initial state (lh1, $\mathbf{k}_{\|}(i) = 0$), to a lower energy final state (hh1, $\mathbf{k}_{\|}(f)$) as a function of $\mathbf{k}_{\|}(f)$, for symmetric [001] QWs, where $\mathbf{k}_{\|}$ is the in-plane quasi-momentum (1). lh (light hole) and hh (heavy hole) label the dominant content of a particular hole QW subband state. The calculations are for zero temperature so it is the zero point oscillations of the phonon field that cause the hole relaxation process. The numerical results show that all three acoustic phonon modes are emitted in the hole relaxation process as compared to the corresponding cb situation where only longitudinal acoustic phonons are emitted [1]. This is a consequence of the s/p symmetry of the cb/vb bandedges and the effect this has on the form of the respective deformation potential interaction. The emission rate versus $\mathbf{k}_{\|}(f)$ is a very strongly peaked function, peaking just before the phonon mode is lost and the peak rate follows the anisotropy of the hh1 final state subband. We explain this behaviour in terms of a simple model for the overlap integral using infinite square well box states for the lh1/hh1 envelope functions, a phonon plane wave factor representing the h-p interaction operator and the application of the energy/momentum conservation laws for the relaxation process. We demonstrate, that the region of the phonon bandstructrure where the emission rate peaks, is the region where phonon anisotropy effects are the strongest. This suggests that phonon anisotropy effects could contribute significantly to the values of h-p emission rates [1].

[1] K. Greipel and U. Rössler, Semicond. Sci. Technol., **5**, 487, (1992)

25. Dynamics of Resonant Tunneling Domains in Superlattices: A Discrete Drift Model, O.M. Bulashenko, L.L. Bonilla, J. Galán, J.A. Cuesta, F.C. Martínez, and J.M. Molera, *Escuela Politécnica Superior, Universidad Carlos III de Madrid, Butarque 15, 28911 Leganés, Spain.* Transport of photoexcited carriers under an applied electric field perpendicular to the layers of an undoped GaAs/AlAs superlattice is described self-consistently. At high excitation laser power, the electric field throughout the superlattice becomes nonuniform by breaking up into domains with different field strength [1]. Domain formation results in time-periodic damped oscillations of the photocurrent (PC) and the photoluminescence (PL) spectrum experimentally observed by S.H. Kwok *et al* (S.H. Kwok, R. Merlin. H.T. Grahn, K. Ploog, L.L. Bonilla, J. Galán, J.A. Cuesta, F.C. Martínez and J.M. Molera, preprint (1994)]. The static I-V curves present then a characteristic oscillatory behavior with the voltage period given approximately by the inter-subband separation [2]. To explain these phenomena we present a discrete drift model whose main ingredients are negative differential resistance due to sequential resonant tunneling (SRT) and the assumption that the state of the superlattice is defined by averages of the electric field and the carrier densities (electron density in the three subbands involved in the SRT and density of holes) over each well. The model explains: (i) the static I-V characteristic through the formation of a domain wall connecting high and low field regions, and (ii) the PC and PL oscillations after the domain wall is formed. In our model, they arise from the combined motion of the wall and the shift of the values of the electric field at the domains. The model predicts quantitatively: (i) the threshold for the photogeneration rate (proportional to the laser power) above which the uniform field distribution breaks into domains. Below this rate the nonuniform field distribution with domains is metastable and it eventually relaxes to the uniform field distribution (after the

PC oscillations have stopped); (ii) characteristic width of the domain wall and its time dependence.

[1] H.T. Grahn *et al.*, Phys. Rev. B 41, 2890 (1990).

[2] L. Esaki and L.L. Chang, Phys. Rev. Lett. 33, 495 (1974).

26. The Dislocation Mechanism of Photoinduced Charge Carrier Transport in Semiconductor Multi-structure Nanocrystal, O.I. Shevaleevskii and M.A. Kojuschner, *N.N. Semanov Institute of Chemical Physics, RAS, Kosigin st. 4, 117977 Moscow, Russia.* The multistructure nanocrystals of some semiconductors with so-called core-shell structure give unexpectedly high photosensitive response of the outer crystalline structure of the particle. The particles under consideration of about 100 nm in diameter have crystal semiconductor core that is covered by a shell crystal layer with different crystalline structure of about 10-20 lattice parameters thick. It was shown that the discrepancy between core and shell lattice parameters give rise to dislocation structure formation in the outer layer of the small particle. According to this we assumed the modification of transport characteristics in nanoscale outer layer of the particle and made the appropriate estimations. As a result we present here the dislocation model of photoinduced charge carrier transport for core-shell nanoscale particle. It was shown that dislocations play dominant role in all the steps of photoinduced charge carrier transport. The kinetic properties of the electron transfer in the vicinity of dislocation and the estimated values of electron mean free pass show the probability of the electron to be captured by the dislocation is high enough and so the main transport parameters of the outer layer should be described as dislocation one. The lifetime of the electron captured by dislocation and its following hopping conduction via dislocation states make this mechanism most efficient to be compared with the bulk one. We also spread our estimations to the systems with mixed electron and ion conductivity. We considered multistructure AgHal particles where the discrepancy of lattice parameters was about 3%. It was shown that the light induced conductivity of Ag ions would be some orders of magnitude higher when coming along dislocation tube. The whole mechanism involving photostimulated process of ion and electron propagation in these particles give rise to Ag cluster formation on the surface of the particle (Mott and Harney). We showed that in multistructure AgHal particles this process comes with the help of dislocation structure and the resulted Ag cluster is forming in the vicinity of surface defect that accomplish the surface end of the dislocation. The whole situation allows us to speak about dislocation mechanism of latter image formation in multistructure AgHal particles.

27. Nonequilibrium Phenomena in Split Gate Quantum Waveguides, A. Ecker and S.M. Goodnick, *Dept. of Electrical and Computer Engr., Oregon State University, Corvallis, OR 97331*, and C. Berven and M.N. Wybourne, *Physics Dept., Univ. of Oregon, Eugene, OR 97403.* While near equilibrium transport through quantum point contacts and split gate quantum waveguides has been well studied, relatively little work has been concerned with far from equilibrium transport under large source-drain bias in these structures. We have previously reported current bistability in quantum dots formed by double constrictions in GaAs/AlGaAs electro-static confinement structures.[1] Here we report a detailed theoretical and experimental study of nonequilibrium transport through a single constriction as a function of bias across the structure (source-drain bias) for various gate bias (barrier height) conditions. For low source-drain bias at 1.5 *K*, the experimental nonlinear multi-mode conductance is found to be in agreement with the predictions of mode-matching theory and simple adiabatic transport based on the generalized Landauer-Büttiker formula. In samples which are illuminated with bandgap radiation in situ, a marked rise in current is evident for all gate biases for source-drain voltages greater than some critical value (usually on the order of 10-20 mV). For large gate biases such that the initial current is small, the transition from low conductance to high conductance is marked by a significant S-type negative differential conductance regime. Some

samples are observed to exhibit several such negative conductance regimes for different combinations of source-drain bias and gate bias. We have generalized the Landauer-Büttiker formula to include Zener breakdown due to inter-mode tunneling, but the calculated effect is too small to account for the apparent breakdown phenomena in the experimental data. The mode-matching model is combined with three dimensional semiclassical simulation to study the realistic potential due to electro-static confinement in the 2DEG. The inclusion of random impurities in the 1D channel is further studied as a source of intermode tunneling under applied bias in order to explain the departure from ideal behavior. However, agreement may be obtained by considering hot carrier heating of electrons injected over the barrier, and the resultant instability in the thermionic emission current through the constriction as was shown for the double constriction case.[2]

[1] J.C. Wu, M.N. Wybourne, C. Berven, S.M. Goodnick and D.D. Smith, Appl. Phys. Lett. **61**, 2425 (1992).

[2] S.M. Goodnick, J.C. Wu, M.N. Wybourne and Doran D. Smith, Phys. Rev. B **48**, 9150 (1993).

Supported by grants from the National Science Foundation No. ECS 9216768 amd ONR.

28. Model and Transport in Three-Layered Heterostructure with Thin Quantum Well in the Schottky Layer, E.V. Buzaneva, V.V. Levandovskiy, *Radiophysical Faculty of Kiev University, Vladimirskaya 64, 254017 Kiev-17, Ukraine.* In recent years multilayered structures containing low-dimensional regions have received considerable attention due to the possibility of creation of devices with new properties. The system considered in our report is three-layered. It consistsf of one layer of narrow gap semiconductor (GaAs) of thickness 3-10 nm grown on AlAs support and followed by 10-50 nm of AlAs MBE layer covered by metallic contact creating Schottky barrier. Two-dimensional region appears in narrow gap layer of small thickness. In this paper we present the results of computer simulation. Exact solution of Schrödinger equation, specific capacitance and I-V characteristics re obtained and analyzed. Unlike well investigated DH HEMT-like systems, holes and electrons are not spatially separated, that allows light absorption and makes such structure attractive for application in fast electro-optical modulators.

29. Conductance Jumps in STM at Room Temperature in Air, V.V. Dremov, S. Yu. Shapoval, and E.V. Sukhorukov, *Institute of Microelectronics, Russian Academy of Sciences, 142432 Chernogolovka, Moscow District, RUSSIA.* Currently there is a great deal of interest in the application of a scanning probe microscopy technique (SPM) for the investigation, testing and formation of nanoelectronics elements. For these purposes it is necessary to pay attention to the properties of nanoscale contacts formed by an interaction between SPM probes and the sample. Earlier the properties of point contacts[1] and junction "probe-sample"[2,3] were studied. However, there are still several questions in understanding the problem, the main of which is the nature of conductance of such small size contacts. Moreover, the possibility of conductance quantization in monoatomic contacts similar to that in microjunctions based on 2DEG[4] has been discussed in the latest works. Using STM[2,3] and the mechanically controlled break junctions[1], the conductance jumping in the units of $2e^2/h$ vs. the distance between two electrodes has been demonstrated. There are several ways to explain this phenomenon. Nevertheless, the main problem remains to understand why the conductance step value is equal to $2e^2/h$ with high accuracy for junctions based on different metals. It is necessary to note that as a rule such kind of experiments are carried out at helium temperature. However, the simplest evaluation allows us to conclude that mechanical and electron transport properties of monoatomic contacts slightly depend on temperature up to room temperature. In our report we present the experimental results which demonstrate the conductance quantization in the micro-junction between tungsten tip and gold film surface at room temperature in air. For tip preparation we used technique described in [5]. This

is based on a combination of electrochemical ac/dc methods for tungsten etching and ECR-plasma treatment to make tip clean and to avoid following contamination. A gold film thermally evaporated on Si with a thickness of about 100 nm was used as the sample. All experiments were carried out with special constructed STM, that has a small drift (about 0.1 nm during 10 min.) and the resolution at least 0.01 nm along z direction. The aim of our experi-ments was to measure the current through the contact as a function of displacement of the specimen with broken STM feedback. To increase signal to noise ratio the current was measured 128 times at every point, then the average value was recorded. The current vs. displacement dependence has stairstep character. The value of step corresponds to conductance jump of $2e^2/h$. Analysis of our results leads to the conclusion that the main reason of such conductance behavior is the atom jumping between the tip and sample surface under van der Waals interaction. In the other side the equality of conduc-tance steps to $2e^2/h$ indicates the wave-guide behavior of electron movement through the microjunction, similar to that in 2DEG.

[1] J.M. Krans, C.J. Muller, and I.K. Yanson, Phys. Rev. B **48**, 14721 (1993).

[2] N. Agrait, O.G. Rodrigo, and S. Vieira, Phys. Rev. B **47**, 12345 (1993).

[3] J.I. Pascual, J. Mendez, and J. Gomez-Herrero, Phys. Rev. Lett. **71**, 1852 (1993).

[4] B.J. van Wees, H. van Houten, C.W.J. Beenakker, and J.G. Willamson, Phys. Rev. Lett. **60**, 848 (1988); D.A. Waram, T.J. Thornton, R. Newbury, and M. Pepper, J. Phys. C **21**, L209 (1988).

[5] V.V. Dremov, V.A. Makarenko, and S. Yu. Shapoval, Nanobiology (UK), to be published.

30. Quantum Transport within the Launcher-Receptor Electron Device, Emmanuel A. Anagnostakis, *Postgraduate Faculty of Electronics and Telecommunications of the Hellenic Army, 22 Kalamakiou Avenue, GR 174 55 Alimos, Athens, Greece.* In the present work a design is proposed of a vertical transport device relying upon resonant quantum mechanical electron tunneling from a charge launcher element (CLE) into a charge receptor pocket (CRP) within an illuminated semiconductor heterostructure, termed the launcher-receptor electron device (LRD). The function of the illumination, achieved through successive exposures of the LRD to regulated photon doses, is to eventually populate the CLE excited sub-level with electrons which, favored by an intended energy matching, subsequently find themselves inhabiting the CRP fundamental sub-level by virtue of resonant quantum tunneling. The CLE is visualized to be a narrow -spike of high sheet density Si donors embedded within an AlGaAs layer of relatively low background Si doping at a location well below its surface but near its contact to a non-intentionally doped, semi-insulating GaAs substrate, hosting an approximately triangular quantum well, as the CRP of the LRD configuration. The plausibility of the proposed design is supported by a realistic simulation of such a LRD heterostructure.

31. Large Plasmon-Mediated Coulomb Frictional Drag in Coupled Quantum Wells, Ben Yu-Kuang Hu and Karsten Flensberg, *Mikroelektronik Centret, Denmark Technical University, DK-2800 Lyngby, Denmark.* When two doped quantum wells are placed in close proximity and current is driven through one of them, the inter-well electron interactions elicit a frictional drag current in the other layer. The frictional Coulomb drag rate (Fermi temperature) experimentally goes as , which agrees with calculations using static screening models [3]. However, static screening models ignore the effect of the collective charge density oscillations (plasmons) on. We find that by incorporating dynamical screening into the calculation, for the plasmons within the coupled well system significantly enhance the inter-well Coulomb interactions, leading to an enhancement in over the static screening results of almost an order of magnitude at (for parameters corresponding to the experiment in Ref. [1]). Experimental signatures of this plasmon enhancement in are discussed.

[1] T.J. Gramila, *et al.*, Phys. Rev. Lett. **66**, 1216 (1991).

[2] U. Sivan, *et al.*, Phys. Rev. Lett. 68, 1196 (1992).
[3] A.P. Jauho and H. Smith, Phys. Rev. B 47, 4420 (1993).

32. Small Signal Differential Mobility of Planar Superlattice Miniband Transport and Negative Differential Conductance, X.L. Lei[1,2], N.J.M. Horing[2], H.L. Cui[2], and K.K. Thornber[3], *[1]State Key Laboratory of Functional Materials for Informatics, Shanghai Institute of Metallurgy, Chinese Academy of Sciences, 865 Chang Ning Road, Shanghai, 200050, China; [2]Department of Physics jand Engineering Physics, Stevens Institute of Technology, Hoboken, New Jersey, 07030; [3]NEC Research Institute, 4 Independence Way, Princeton, New Jersey, 08540.* We have examined the frequency dependence of carrier drift velocity in a laterally unconfined planar superlattice experiencing miniband conduction in response to a small-signal a.c. electric field superposed on a d.c. bias. Anomalous results quite unexpected from the traditional Drude formula are found in the frequency dependent differential mobility at finite d.c. bias. The real part of the differential mobility as a function of frequency exhibits a broad hump before finally approaching zero at high frequency, and the imaginary part experiences a marked dip to negative values before going through the usual maximum. These effects are due to an increasing rate of momentum relaxation and a decreasing rate of energy relaxation as the bias field is increased, and occur for bias fields which result in positive as well as nega-tive differential mobilities at low frequency. Complex as the various scattering mechanisms are, the frequency dependences can be closely fit with a dual-relaxation time model representing the low-frequency mobility, momentum and energy scattering rates, and the inertial response. Numerical calculations show that the negative-differential mobility persists up to a mobility-transition frequency of order 100 GHz (substantially below the Bloch frequency, which is usually in the THz regime), above which the differential mobility becomes positive.

33. A First Step for Semiconductor Quantum Device Modeling with Incoherent Scattering, Roger Lake, *Corporate R&D, Texas Instruments, Inc., Dallas, TX 75265.* Despite intense experimental investigation of semiconductor quantum devicesf over the last decade[1], there has been comparatively little progress in high-bias quantum device modeling and few reports of quantum device simulators which include, in some approximation, the effects of incoherent scattering from the random potentials of either phonons, impurities, interface roughness, alloy disorder, or other electrons. The methods, Wigner function [2,3,4], non-equilibrium Green's function [5,6], single-electron scattering theory [7,8], and single-electron Schrödinger equation [9], that have been applied to the problem, are varied. The points of view of the different methods can be roughly divided into two categories: non-equilibrium, quantum, statistical physics and single-electron scattering theory. Until recently, there was little consensus on the fundamental approach to the problem. The system to be modeled is a small quantum system containing many electrons interacting with many phonons, impurities, other electrons, etc. The system is connected to reservoirs of different electrochemical potentials. Such a system is a non-equilibrium, quantum statistical system and modeling the system is a problem of non-equilibrium quantum statistical mechanics. I do not imply that the single-electron scattering picture is wrong. It is adequate for many applications. Non-equilibrium statistical mechanics is the fundamental starting point from which we can evaluate the approximations implicit in other approaches. The first two paragraphs written by Langreth in this series in 1976 are still relevant today [10]. I believe that the time is ripe for the application of the non-equilibrium perturbation theoretic method, i.e. the non-equilibrium Green's function theory, to semiconductor quantum device modeling. There have been only two modest efforts at writing general purpose semiconductor device simulators based on the non-equilibrium Green's function approach [5,6]. Both works included only simplified models of the electron-phonon

interaction, but treatedf the interaction in the self-consistent Born approximation. Below, I describe an approach that includes realistic models of the various interactions but only in the first Born approximation. The approach is numerically tractable, easy to implement, and provides a reasonable treatment of scattering for one of the more popular quantum devices, the resonant tunneling diode. Furthermore, it is the first iteration of a self-consistent Born calculation, so it provides a foundation for a self-consistent treatment of the interactions. Since the approach is non-iterative (except for a calculation of the Hartree potential) and relies only on knowledge of the bare Green function, I believe that the approach can be used to incorporate scattering into more sophisticated, and therefore more numerically intensive, treatments of band structure.

[1] My focus is on quantum devices that can be operated at room temperature under high bias such as resonant tunnel-ing diodes (Resonant Tunneling in Semi-conductors: Physics and Applications, edited by L.L. Chang, E.E. Mendez, and C. Tejedor, Plenum, New York, (1991) and hot-electron transistors (S. Luryi in Heterojunction Band Discontinuities Physics and Device Applications, edited by F. Capasso and G. Margaritonodo, p. 489, North-Holland, New York, (1987).

[2] U. Ravaioli, M.A. Osman, W. Poetz, N. Kluksdahl, and D.K. Ferry, Physica, **134B**, 36, (1985).

[3] N.C. Kluksdahl, A.M. Kriman, D.K. Ferry, and C. Ringhofer, Phys.Rev. B, **39**, 7720, (1989).

[4] W.R. Frensley, Rev. Mod. Phys., **62**, 745 (1990) and references therein.

[5] E.V. Anda and F. Flores, J. Phys.: Condens. Matt., **3**, 9087 (1991).

[6] R. Lake and S. Datta, Phys. Rev. B, **45**, 6 670 (1992).

[7] N.S. Wingreen, K.W. Jacobsen, and J.W. Wilkins, Phys. Rev. Lett., **61**, 1396 (1988); Phys. Rev. B. **40**, 11834 (1989).

[8] N. Zou and K.A. Chao, Phys. Rev. Lett., **69**, 3224 (1992).

[9] P. Roblin and W. Liou, Phys. Rev. B, **40**, 2146 (1993).

[10] D.C. Langreth in 1975 NATO Advanced Study Institute on Linear and Nonlinear Electron Transport in Solids, p. 3, (Plenum Press, New York, 1976).

34. Theory of Delta-Wires, A. Shik, *Ioffe Physical-Technical Institute, 194021 St.-Petersburg, Russia.* High-quality vicinal surfaces of semiconductors are known to contain periodic system of monolayer steps. If we grow a sub-monolayer of doping impurity on this surface, impurity atoms will stick mainly to these steps and at first growth stages will form separate linear chains. If the linear density of impurities in such chain N exceeds the inverse effective Bohr radius , impurities are ionized and carriers can move along the chain. In other directions their motion is restricted by the attractive impurity potential and energy spectrum is quantized. So, the structure considered is a new type of quasi-one-dimensional electron systems. By analogy with -layers, we call these structures -wires. To calculate the main properties of delta-wires, we consider impurity charge as a homogeneously charged line. To calculate the resulting self-consistent potential (p is the distance from the wire axis) and energy levels in this potential E_{NM}, we must solve self-consistently the system of Poisson and Schrödinger equatins. The system contains only one dimensionless parameter $\nu N a_B$ and such characteristics as the number of discrete energy levels, their particular values E_{NM}, populations of corresponding one-dimensional subbands n_{NM}, etc. in dimensionless units are presented as universal functions of ν, by analogy with the -layer theory [1]. Kinetic properties of -wires have been also calculated. At not very high temperatures electron mobility is governed by scattering at the same linear impurity chain which forms the wire itself. For a typical situation of more than one occupated subbands, both intra- and inter-subband scattering is to be taken into account. Since the impurity concentration in the absence of compensation is equal to the electron concentration and the scattering is due to Coulomb forces, it is determined by the same parameter ν. As a result, the partial mobilities in subbands and the wire

conductance are real so presented as universal functions of *v*.

[1] O.A. Mezrin and A. Shik, Superlatt. & Microstr. **10**, 107 (1991).

35. Electron Heating in GaAs due to Electron-Electron Interactions, B. Brill and M. Heiblum, *Braun Center for Sub Micron Research, Dept. of Condensed Matter Physics, The Weizmann Institute of Science, Rehovot 76100, Israel.* The interactions of hot electrons in semiconductor layers were usually studied assuming the scattering mechanisms to affect only the injected hot electron distribution, leaving the cold electrons and the lattice in thermal equilibrium. However, if the energy transfer from hot to cold electrons is sufficiently large, the cold electron distribution can be expected to be significantly modified. We have studied the interactions of injected hot electrons, passing briefly through a thin doped GaAs layer, with cold electrons confined in the layer. This situation allows us to discuss separately the cold and hot electron distributions, and our device allows us to probe both. The study was carried out using a Tunneling Hot Electron Transfer Amplifier (THETA), with a heavily doped base, confined between two barriers, serving as the transport layer. In this device, hot electrons are injected by tunneling through a thin AlGaAs barrier (emitter barrier) into the base and are collected over a second AlGaAs barrier (collector barrier). To surmount the collector barrier the collected electrons must have a normal energy higher than the barrier height, . In order to probe the heated cold electron distribution the collector barrier is made very low. The differential transfer ratio, where is the injected current and is the collected current, measures the probability for electrons, injected within a narrow energy range into the base, to be collected over the collector barrier. Surprisingly, at high injection energies we measure values much above unity, i.e., the output current is larger than the injected current, indicating that a current amplification process is taking place in the layer. Two possible pictures were considered as possible explanations: one assuming single particle energy and momentum transfer from hot to cold electrons; and the other assuming that thermal equilibrium is reached among the heated cold electrons, with an elevated electron temperature, higher than the lattice temperature, leading to thermionic emission over the low collector barrier. We discriminate between the two pictures by noting that in the single particle picture the heating effect depends on the injected electron energy only whereas in the equilibrium picture the injected power, depending on both the injection energy and current, is the important parameter. Comparing devices which differ by only the emitter barrier thickness, allowing thus measure-ments with the same injection energy but different currents, we found that the heating effect clearly depends on the injected power rather than the injection energy alone. We thus conclude that the equilibrium picture is more relevant to the experimental situation. Two complementary methods were used to measure the electron temperature in the base. One method utilizes the fact that thermionic emission over the low collector barrier results in a partial shunting of the base to the collector, thus reducing the (ac) base resistance. The other method utilizes the effect of weak localization on the magnetoresistance, which is a sensitive function of temperature. The two methods were calibrated by heating the lattice without hot electron injection and then used when the lattice was kept at 1.4 K and hot electrons were injected. We found the temperature to be 10 - 20 K. monotonically dependent on the injected power. At the lower injection energies the two measurements agreed well, however at higher injection energies the base resistance measurement showed a higher temperature, reflecting the existence of temperature gradients along the base layer. The measured temperatures are consistent with the existence of a significant thermionic current, leading to . Adopting the equilibrium picture and using the relaxation time approximation we express the condition for steady state in the base as an energy balance equation. Using the measured temperatures and adopting the cooling rate due to acoustical phonons from the literature the energy relaxation time for the hot injected electrons can be extracted. The relaxation time found for electrons

injected at an energy of 130 meV above the Fermi see of a doped layer with n=2 $10^{18}cm^{-3}$ was found to be 250±100 fs. which is two times smaller than published theoretical times calculated for similar conditions.

36. Quantum Hydrodynamics: Derivation and Classical Limit, I. Gasser and P.A. Markowich, *Fachbereich Mathematik, TU-Berlin, Straße des 17, Juni 136, W-10623, Berlin, Germany.* Recently many simulations of quantum effects in semi-conductor devices using the quantum hydrodynamic model (QHD) were presented [1]. The simulations showed that the QHD is a promising possibility to model ultraintegrated devices. The QHD can be obtained from the classical hydrodynamic model by adding quantum correction terms of order $\bar{h}^2$ ($\bar{h}$ denotes the scaled planck constant). We are interested both in a rigorous justification of the model and the classical limit $\bar{h}\to 0$, e.g. if and in which sense the solutions of the classical hydrodynamic model approximate the solutions of the QHD for small values of $\bar{h}$. Introducing the wave function, the particle density, the particle current density, it can be seen easily that the nonlinear Schrödinger equation is (formally) equivalent to the QHD equations. Obviously, it is not possible to obtain an energy (or temperature) equation from a single Schrödinger equation. In the case of constant pressure we are able to perform the classical limit rigorously, passing from the Schrödinger equation (wave function approach) to the Wigner equation, which is the quantum equivalent to the classical Vlasov equation. Euler type equations are not equal to the formal limit of the above QHD equations with constant pressure.

[1] C.L. Gardner, Resonant Tunneling in the Quantum Hydrodynamic Model, VLSI Design, to appear 1994.

[2] I. Gasser and P.A. Markowich, Quantum Hydrodynamics, Wigner Transforms and the Classical Limit, submitted 1994.